# Legal Knowledge and Information Systems

**JURIX 2025: The Thirty-eighth Annual Conference, Turin, Italy, 9-11 December 2025**

**Editors:**
Réka Markovich, Luigi Di Caro, Amon Rapp and Claudio Schifanella

**JURIX 2025**

IOS Press

# LEGAL KNOWLEDGE AND INFORMATION SYSTEMS

# Frontiers in Artificial Intelligence and Applications

The book series Frontiers in Artificial Intelligence and Applications (FAIA) covers all aspects of theoretical and applied Artificial Intelligence research in the form of monographs, selected doctoral dissertations, handbooks and proceedings volumes. The FAIA series contains several sub-series, including 'Information Modelling and Knowledge Bases' and 'Knowledge-Based Intelligent Engineering Systems'. It also includes the biennial European Conference on Artificial Intelligence (ECAI) proceedings volumes, and other EurAI (European Association for Artificial Intelligence, formerly ECCAI) sponsored publications. The series has become a highly visible platform for the publication and dissemination of original research in this field. Volumes are selected for inclusion by an international editorial board of well-known scholars in the field of AI. All contributions to the volumes in the series have been peer reviewed.
All volumes within the FAIA series are indexed in ACM Digital Library; DBLP; EI Compendex; Google Scholar; Scopus; Zentralblatt MATH. All volumes are submitted for individual evaluation by Web of Science: Conference Proceedings Citation Index (CPCI).

Series Editors:
Nicola Guarino, Pascal Hitzler, Joost N. Kok, Jiming Liu, Ramon López de Mántaras,
Riichiro Mizoguchi, Mark Musen, Sankar K. Pal, Ning Zhong

## Volume 416

*Recently published in this series*

Vol. 415. E. Cortés-Ancos, D.A. López-García and A.J. Tallón-Ballesteros (Eds.), Electronics, Communications and Networks – Proceedings of the 15th International Conference (CECNet 2025), Hong Kong, 24-26 November 2025
Vol. 414. V. Piuri, D. Pelusi, M. Paprzycki, A. Ghosh and P. Coscia (Eds.), Advancements in Smart Innovations, Intelligent Systems, and Technologies – Proceedings of the 1st International Conference on Innovations in Engineering and Emerging Technologies (ICIEET 2025), Biel/Bienne, Switzerland, 9-10 June 2025
Vol. 413. I. Lynce, N. Murano, M. Vallati, S. Villata, F. Chesani, M. Milano, A. Omicini and M. Dastani (Eds.), ECAI 2025 – 28th European Conference on Artificial Intelligence, 25-30 October 2025, Bologna, Italy
Vol. 412. A. Lau and A.J. Tallón-Ballesteros (Eds.), Digitalization and Management Innovation IV – Proceedings of DMI 2025, Hong Kong, 7-9 July 2025
Vol. 411. H. Fujita, A. Hernandez-Matamoros and Y. Watanobe (Eds.), New Trends in Intelligent Software Methodologies, Tools and Techniques – Proceedings of the 24th International Conference (SoMeT_25), Kitakyushu, Japan, 23-26 September 2025
Vol. 410. K. Trejo, I. Aguiló, J.V. Riera and J. Pascual (Eds.), Artificial Intelligence Research and Development – Proceedings of the 27th International Conference of the Catalan Association for Artificial Intelligence

ISSN 0922-6389 (print)
ISSN 1879-8314 (online)

# Legal Knowledge and Information Systems

JURIX 2025: The Thirty-eighth Annual Conference, Turin, Italy, 9-11 December 2025

Edited by

Réka Markovich

*University of Luxembourg, Luxembourg*

Luigi Di Caro

*University of Turin, Italy*

Amon Rapp

*University of Turin, Italy*

and

Claudio Schifanella

*University of Turin, Italy*

Sage | IOS Press

© 2025 The Authors.

This book is published online with Open Access and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).

ISBN 978-1-64368-638-7 (online)
doi: 10.3233/FAIA416



IOS Press

A Sage company
Teleportboulevard 120
1043 EJ Amsterdam

Sage
1 Oliver's Yard
55 City Road
London EC1Y 1SP

Sage
2455 Teller Road
Thousand Oaks
California 91320

Sage
Unit No 323-333, Third Floor, F-Block
International Trade Tower
Nehru Place, New Delhi – 110 019

Sage
8 Marina View Suite 43-053
Asia Square Tower 1
Singapore 018960

DISCLAIMER
The authors, editors, and publisher will not accept any legal responsibility for any errors or omissions that may be made in this publication. The publisher makes no warranty, express or implied, with respect to the material contained herein.

# Foreword

Interest in technological advances – be that processing, or representation and engineering – as they relate to legal information and knowledge, has been growing for decades. Formal and computational models of legal reasoning, legal data analytics, legal information retrieval, the application of machine learning techniques to various legal tasks, and experimental evaluations of these systems have all been active research areas, and seem likely to continue to be so.

This book presents the proceedings of JURIX 2025, the 38th International Conference on Legal Knowledge and Information Systems, held from 9 to 11 December in Turin, Italy, under the auspices of the JURIX Foundation for Legal Knowledge-Based Systems, and hosted by the University of Turin. The annual JURIX conference has become an international forum for academics and professionals to exchange knowledge and experiences at the intersection of law and artificial intelligence (AI). For this edition, we received 139 eligible submissions (of which 103 were long papers) by 392 authors from 35 countries. Following a rigorous review process, carried out by a programme committee of 107 experts recognised in the field, 22 submissions were selected for publication as long papers and 18 as short papers, representing a 21.3% acceptance rate for long papers (per long submissions, 15,8% for long papers per all submissions) and 28.7% overall (accepted for presentation/all submissions). An additional 18 submissions were accepted as posters. The wide range of topics covered includes formal approaches (deontic logic, cased-based reasoning, formal argumentation, and other formalisms) applied to various aspects of legal reasoning, machine learning (large language models, including conversational agents) and information retrieval methods applied to various natural language processing tasks, hybrid approaches to working on the frontiers between symbolic and sub-symbolic methods, experimental inquiries on interfaces between computational systems and legal systems, and network analysis in the law.

Providing an overview of recent advances, the book will be of interest to all those working at the intersection of law and AI.

# Preface JURIX 2025

We are pleased to present the proceedings of JURIX 2025, the 38th International Conference on Legal Knowledge and Information Systems. Organised under the auspices of the Foundation for Legal Knowledge-Based Systems (https://www.jurix.nl), the JURIX annual conference has become established as an internationally renowned forum for the exchange of ideas concerning theoretical models and practical applications in the broadly construed domain of artificial intelligence (AI) and law research, including legal information systems and legal knowledge systems research. Traditionally, the field of AI and law has been concerned with legal knowledge representation and engineering, logic, and computational models of legal reasoning, legal data analytics, and legal information retrieval, but recent years have witnessed the rise to prominence of the application of machine learning tools to legally relevant tasks, recently joined by a sharp increase in the exploitation of large language model (LLM)-based AI agents and systems for law-related tasks. Furthermore, the constantly growing influence of AI on different spheres of social life has prompted interest in the explainability, trustworthiness, and responsibility of computational systems. From their inception, the JURIX conferences have been a venue for interdisciplinary research, integrating methods, approaches, and conceptual frameworks from different branches of computer science and jurisprudence, including cognitive and socio-technical dimensions.

The 2025 edition of JURIX, which runs from 9 to 11 December, is hosted by the University of Turin, in Turin, Italy. An exponential growth in the number of submissions was experienced for this edition compared to previous years, with 139 eligible submissions (of which 103 were long papers) by 392 authors from 35 countries being received. Following a rigorous review process, carried out by a programme committee of 107 experts recognised in the field, 22 submissions were selected for publication as long papers and 18 as short papers, representing a 21.3% acceptance rate for long papers (per long submissions, 15.8% for long papers per all submissions) and 28.7% overall (accepted for presentation/all submissions). An additional 18 submissions were accepted as posters. The accepted papers cover a wide range of topics, including formal approaches (deontic logic, case-based reasoning, formal argumentation, and other formalisms) applied to various aspects of legal reasoning, machine learning, natural language processing and information retrieval methods as applied to various legal tasks, hybrid approaches working on the frontiers between symbolic and sub-symbolic methods, experimental enquiries on interfaces between computational systems and legal systems, and network analysis in law.

Two invited speakers, Marie-Francine Moens and Tiago Cabral, have honoured JURIX 2025 by kindly agreeing to deliver keynote lectures.

Marie-Francine Moens pioneered in NLP for AI&Law; we now live in a world (partly) shaped by her. She is currently a professor emerita at KU Leuven, Belgium. From 1992 to 2007, she was a researcher, guest professor and associate professor at the Centre for Law and ICT, and later became a professor and full professor at the Department of Computer Science at KU Leuven.

Marie-Francine Moens received the ERC Advanced Grant CALCULUS (2018-2024), awarded by the European Research Council. From 2012 to 2016, she was the coordinator of the MUSE project financed by Future and Emerging Technologies (FET) of the European Commission. In 2021, she was the general chair of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021). In 2011 and 2012, she was appointed as chair of the European Chapter of the Association for Computational Linguistics (EACL) and was a member of the executive board of the Association for Computational Linguistics (ACL). She is currently associate editor of the journal IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) and of the journal AI Open. In 2014, she was a Scottish Informatics and Computer Science Alliance (SICSA) Distinguished Visiting Fellow. She is a fellow of the European Laboratory for Learning and Intelligent Systems (ELLIS). Her research topics include machine learning for natural language processing; representation learning for language grounding in the physical and social world; deep learning, latent variable models and brain-inspired models; and machine learning models for structured prediction and generation.

Tiago Sérgio Cabral is a Legal and Policy Officer at the European Commission, where he works in the European AI Office in the team responsible for AI Regulation and Compliance. He specializes in Artificial Intelligence, Data Protection, and Cybersecurity. He holds a PhD in Legal Sciences – Specialization in Public Legal Sciences, a Master's in European Union law, and an LL.B. Before joining the European Commission, Tiago worked as a lawyer in various international law firms. He is also frequently involved in research projects and lecturing commitments.

This year, continuing a tradition, JURIX has also been accompanied by satellite co-located events: five workshops – CLAIRvoyantS, AI4Legs-V, NLL2FR2025, AI4EVIR, and AI4A2J – and the Doctoral Consortium.

*CLAIRvoyantS: 3° ConventicLE on Artificial Intelligence Regulation and Safety* explored topics such as the impact of the AI Act on high-risk and sensitive sectors (e.g. the judicial, legal, and health domains), and questions whether AI tools are currently transparent, accurate, and trustworthy enough.

*AI4Legs-V: The 5th Workshop on AI for Legislation* discussed challenging questions involving the use of AI and technology in general to support the legislative process, with interdisciplinary instruments coming from the philosophy of law, constitutional law, legal informatics, including AI and law, computational linguistics, computer science, HCI, and legal design.

*NLL2FR2025: Translating Natural Legal Language into Formal Representation* brought together researchers and practitioners interested in bridging the gap between natural legal language and formal representations, but the focus of the workshop extended beyond the translation of natural language rules into logical formulae to include the formalisation of legal cases described in natural language.

*AI4EVIR: AI for Evidential Reasoning* aimed to bring together researchers working on formal, computational, and empirical approaches to reasoning with evidence, as well as those with expertise in forensic science and evidence evaluation. The goal of the workshop was to foster discussion and exchange between theoretical and applied perspectives on how AI can contribute to evidential reasoning in legal and investigative contexts.

*AI4A2J: AI for Access to Justice, Dispute Resolution, and Data Access* invited legal technologists, researchers, and practitioners to discuss how innovations in AI can help to close the access to justice gap: the majority of legal problems go unresolved

around the world because potential litigants lack the time, money, or ability to participate in the court processes which could solve their problems. The workshop had three main tracks: i) Data issues related to access to justice (building reusable, shareable datasets for research), ii) AI for access to justice generally, and iii) AI for dispute resolution.

Organising this edition would not have been possible without the support of many people and institutions. Special thanks are due to Francesca Grasso (Local Chair), Rachele Mignone (Web and Social Media Chair), Ivan Spada and Federico Torrielli (Tech Chairs), and the local support team of the University of Turin, Davide Audrito, Marinella Quaranta, Ilaria Angela Amentea, Marianna Molinari, and Stefano Locci. We would like to thank the workshop organisers for their proposals and for the effort involved in organising the events. We owe our gratitude to Monica Palmirani, who kindly assumed the function of Doctoral Consortium Chair this year as well. We are particularly grateful to the 107 members of the Programme Committee for their work in the rigorous review process and for their participation in the discussions.

Finally, we would like to thank Jaromír Savelka and Michal Araszkiewicz for their advice, and the current JURIX General Board for their support and for taking care of all JURIX initiatives.

Réka Markovich, JURIX 2025 Programme Chair
Luigi Di Caro, JURIX 2025 Organising Co-Chair
Amon Rapp, JURIX 2025 Organising Co-Chair
Claudio Schifanella, JURIX 2025 Organising Co-Chair

# About the Conference

**Program Chair**
Réka Markovich, University of Luxembourg, Luxembourg

**Organizing Chairs**
Luigi Di Caro, University of Turin, Italy
Amon Rapp, University of Turin, Italy
Claudio Schifanella, University of Turin, Italy

**Program Committee**
Fahimeh Abedi, University of Melbourne, Australia
Tommaso Agnoloni, ITTIG-CNR, Italy
Thomas Ågotnes, University of Bergen, Norway
Michał Araszkiewicz, Jagiellonian University, Poland
Joaquín Arias, Universidad Rey Juan Carlos, Spain
Kevin Ashley, University of Pittsburgh, United States
Katie Atkinson, University of Liverpool, United Kingdom
Davide Audrito, University of Turin, Italy
David Bareham, University of Liverpool, United Kingdom
Nick Bassiliades, Aristotle University of Thessaloniki, Greece
Floris Bex, Utrecht University, The Netherlands
Ilaria Canavotto, University of Maryland, United States
Pompeu Casanovas, Universitat Autònoma de Barcelona, Spain
Marcello Ceci, University of Luxembourg, Luxembourg
Damien Charlotin, HEC Paris, France
Jack G. Conrad, Thomson Reuters, United States
Giuseppe Contissa, University of Bologna, Italy
Federico Costantini, University of Udine, Italy
Marina De Vos, University of Bath, United Kingdom
Luigi Di Caro, University of Turin, Italy
Huimin Dong, TU Wien, Austria
Massimo Durante, University of Turin, Italy
Arthur Dyevre, KU Leuven Centre for Empirical Legal Studies, Belgium
Jenny Eriksson Lundström, Uppsala University, Sweden
Mattia Falduti, Free University of Bozen-Bolzano, Italy
Marcelo Finger, University of São Paulo, Brazil
Nicoletta Fornara, Università della Svizzera italiana, Switzerland
Enrico Francesconi, IGSG-CNR, Italy
Kripabandhu Ghosh, Indian Institute of Science Education and Research, India
Marco Giacalone, Vrije Universiteit Brussel, Belgium
Randy Goebel, University of Alberta, Canada
Guido Governatori, Central Queensland University, Australia
Matthias Grabmair, Technical University of Munich, Germany
Margaret Hagan, Stanford Legal Design Lab, United States

Jakub Harašta, Masaryk University, Czech Republic
Mustafa Hashmi, Data 61, CSIRO, Australia
John Horty, University of Maryland, United States
Joris Hulstijn, Utrecht University, The Netherlands
Diana Inkpen, University of Ottawa, Canada
Jeroen Keppens, King's College London, United Kingdom
Mi-Young Kim, University of Alberta, Canada
Francesca Lagioia, University of Bologna, Italy
Marc Lauritsen, Capstone Practice Systems, United States
Sarah Lawsky, University of Illinois, United States
Jochen L. Leidner, Coburg University of Applied Sciences, Germany
Tomer Libal, Enidia AI, Luxembourg
Rūta Liepiņa, University of Bologna, Italy
Davide Liga, University of Luxembourg, Luxembourg
How Khang Lim, Singapore Management University, Singapore
Daniel W. Linna Jr, Northwestern University, United States
Arno R. Lodder, Vrije Universiteit Amsterdam, The Netherlands
Emiliano Lorini, IRIT, France
Juliano Maranhao, University of São Paulo, Brazil
Florian Matthes, Technical University of Munich, Germany
Thorne McCarty, Rutgers University, United States
Jack Mumford, University of Liverpool, United Kingdom
Le-Minh Nguyen, Japan Advanced Institute of Science and Technology, Japan
Paulo Novais, University of Minho, Portugal
Tereza Novotná, Masaryk University, Czech Republic
Gordon Pace, University of Malta, Malta
Monica Palmirani, University of Bologna, Italy
Luca Pasetto, University of Luxembourg, Luxembourg
Ginevra Peruginelli, ITTIG-CNR, Italy
Wim Peters, University of Aberdeen, United Kingdom
Radim Polčák, Masaryk University, Czech Republic
Prakash Poudyal, Kathmandu University, Nepal
Henry Prakken, Utrecht University, The Netherlands
Paulo Quaresma, Universidade de Evora, Portugal
David Restrepo Amariles, HEC Paris, France
Niccolo Ridi, King's College London, United Kingdom
Livio Robaldo, University of Swansea, United Kingdom
Víctor Rodríguez Doncel, Universidad Politécnica de Madrid, Spain
Antoni Roig, Autonomous University of Barcelona, Spain
Arianna Rossi, LIDER Lab, DIRPOLIS, Sant'Anna School of Advanced Studies, Italy
Antonino Rotolo, University of Bologna, Italy
Ken Satoh, Center for Juris-Informatics, ROIS, Japan
Giovanni Sartor, University of Bologna, Italy
Jaromír Savelka, Carnegie Mellon University, United States
Burkhard Schafer, The University of Edinburgh, United Kingdom
Frank Schilder, Thomson Reuters, United States
Marijn Schraagen, Utrecht University, The Netherlands
Erich Schweighofer, University of Vienna, Austria
Giovanni Sileno, University of Amsterdam, The Netherlands

Jerrold Soh, Singapore Management University, Singapore
Christoph Sorge, Saarland University, Germany
Clara Smith, Universidad Nacional de La Plata and Universidad Católica de La Plata, Argentina
Quinten Steenhuis, Suffolk University, United States
Sarah Sutherland, Parallax Information Consulting, United States
István Szakadát, Budapest University of Technology and Economics, Hungary
Dimitrios Tsarapatsanis, University of York, United Kingdom
Leon van der Torre, University of Luxembourg, Luxembourg
Gijs van Dijck, Maastricht University, The Netherlands
Tom Van Engers, University of Amsterdam, The Netherlands
Marc van Opijnen, KOOP, The Netherlands
Wijnand van Woerkom, Max Planck Institute, Germany
Bart Verheij, University of Groningen, The Netherlands
Fabio Vitali, University of Bologna, Italy
Vern Walker, Hofstra University, United States
Bernhard Waltl, Technical University of Munich, Germany
Yueh-Hsuan Weng, Tohoku University, Japan
Hannes Westermann, Maastricht University, The Netherlands
Radboud Winkels, University of Amsterdam, The Netherlands
Adam Wyner, Swansea University, United Kingdom
Minghui Xiong, Zhejiang University, China
Liuwen Yu, University of Luxembourg, Luxembourg
John Zeleznikow, Victoria University, Australia
Tomasz Zurek, Maria Curie-Skłodowska University, Poland

**JURIX General Board**
Katie Atkinson, University of Liverpool, United Kingdom (president)
Michał Araszkiewicz, Jagiellonian University, Poland (vice-president/secretary)
Giovanni Sileno, University of Amsterdam, Netherlands (treasurer)
Enrico Francesconi, IGSG-CNR, Italy
Réka Markovich, University of Luxembourg, Luxembourg

# Contents

Foreword v

Preface vi
*Réka Markovich, Luigi Di Caro, Amon Rapp and Claudio Schifanella*

About the Conference ix

**Long Papers**

Precedent as Interpretive Resource: A Case Frame Analysis of Statutory Reasoning 2
*Michał Araszkiewicz*

JuriDiscovery: Automated Analysis of Legal Dispute Triggered by Public Policies Using Artificial Intelligence. A Case Study Applied to the Individual Microentrepreneur 14
*Loreane Brandizzi and Guilherme Rodrigues*

Unconstitutionality Prevention in Bills Using Hybrid AI 26
*Michele Corazza, Pier Francesco Bresciani, Generoso Longo, Faria Ferooz, Salvatore Sapienza and Monica Palmirani*

Automated Creation of a Legal Knowledge Graph Addressing Cases of Violence Against Women. Resource Methodology and Lessons Learned 37
*Claudia D'Amato, Giuseppe Rubini, Francesco Didio, Donato Francioso, Fatima Zahra Amara and Nicola Fanizzi*

Explaining Discretionary Judicial Decisions Using Answer Set Programming 49
*Josephine Dik*

Contrary-to-Duty Rights: From Hohfeld to Agreement Revision 61
*Huimin Dong, Leendert van der Torre and Liuwen Yu*

Does ChatGPT Understand the Law? A Case Study on Road Homicide in Italy 74
*Grazia Garzo and Alessandro Palumbo*

A Two-Dimensional Evaluation Framework for Factual and Reasoning Assessment of LLMs in Legal Question Answering 86
*Sinan Gultekin, Matteo Rossi Reich, Francesca Galloni, Francesca Lagioia, Elena Consiglio, Giovanni Sartor and Sara Bagnato*

Can LLMs Create Legally Relevant Summaries and Analyses of Videos? 98
*Lyra Hoeben-Kuil, Gijs van Dijck, Jaromir Savelka, Johanna Gunawan, Konrad Kollnig, Marta Kołacz, Mindy Duffourc, Shashank M. Chakravarthy and Hannes Westermann*

Which Neurons Nudge Normative Stance? Causal Tests and Mechanistic Evidence via Contrastive Last-Token Steering 110
*Davide Liga and Liuwen Yu*

Argumentative LLMs for Legal Information Entailment 121
*Charlie Lindsay, Francesca Toni, Tien Dat Nguyen, Tan Minh Nguyen, Ngoc Anh Trang Pham, Quang Huy Chu, Hoang Nhat Nguyen and Le-Minh Nguyen*

The Result Model Under Inconsistent Knowledge: Theory and Experiments 133
*Yoann Morello, Agata Ciabattoni and Morgan Gray*

Context-Aware Citation Networks: A Human–AI Dataset, Analysis, and Tool 145
*Jack Mumford, Francesco Florimonte, Katie Atkinson and Kanstantsin Dzehtsiarou*

PersonaMatrix: A Recipe for Persona-Aware Evaluation of Legal Summarization 156
*Tsz Fung Pang, Maryam Berijanian, Thomas Orth, Breanna Shi and Charlotte S. Alexander*

Machine Learning to Promote Self-Composition in Conflict Resolution: The Brazilian Concil-IA Project 168
*Isabela Sabo, Thiago Dal Pont, Aires Rover and Jomi Hubner*

Validity Diagrams 180
*Diogo Sasdelli and Bianca Steffes*

That's So FETCH: Fashioning Ensemble Techniques for LLM Classification in Civil Legal Intake and Referral 192
*Quinten Steenhuis*

How Anonymous Is Anonymous? A Techno-Legal Exploration 204
*Stephanie von Maltzan, Daniel Vogel, Marc Ohm and Florian Idelberger*

Label-Aware Contextual Retrieval for Few-Shot Legal Argument Mining 216
*Huihui Xu and Kevin Ashley*

A Multi-Agent AI Framework for Legal Aid: Real-World Interactive Dialogue in Case Consultations 228
*Ying-chu Yu, Sieh-Chuen Huang and Hsuan-Lei Shao*

Do LLMs Truly "Understand" when a Precedent Is Overruled? 239
*Li Zhang, Jaromir Savelka and Kevin Ashley*

Multilingual Legislative Definitions Retrieval and Generation Using LLM and Agentic AI 251
*Leonardo Zilli, Michele Corazza, Monica Palmirani and Salvatore Sapienza*

**Short Papers**

A Systematic Approach to Assess Languages Proposed for Rules as Code 264
*Sofia Ali, Giovanni Sileno and Tom van Engers*

Accountability as a Software Engineering Tool for Multi-Agent Systems 270
*Matteo Baldoni, Cristina Baroglio, Roberto Micalizio and Stefano Tedeschi*

Segmentation and Processing of German Court Decisions from Open Legal Data 276
*Harshil Darji, Martin Heckelmann, Christina Kratsch and Gerard de Melo*

An Ontology-Driven Graph RAG for Legal Norms: A Hierarchical, Temporal, and Deterministic Approach 282
*Hudson de Martim*

From Data to Equity: Predicting and Auditing Fairness in Indian Bail Decisions 288
*Sneha Deshmukh and Prathmesh Kamble*

Automated Extraction of Judicial Interpretative Formulas in EU Case Law on VAT 294
*Giulia Grundler, Piera Santin, Alessia Fidelangeli, Rachele Mignone, Federico Galli, Andrea Galassi, Giuseppe Contissa, Luigi di Caro and Paolo Torroni*

A Self-Supervised Method for Legal Reference Resolution 300
*Šimon Horvát, Dávid Varga, Peter Gurský, Nicol Fedurcová, Samuel Baran and Zoltán Szoplák*

Learning from the Judicial Journey: A Predictive Model for Case Disputability 306
*Ho-Chien Huang and Chao-Lin Liu*

Reinforcement Learning with Argument-Structured Reward for Court Decision Abstractive Summarization 312
*Yuntao Kong, Ye Xiong, Shuyuan Zheng and Ken Satoh*

A Hohfeldian Knowledge Base for LLM-Assisted Legal Information Retrieval in Marine Biodiversity Law 318
*Rohan Nanda, Henrique Marcos, Hannes Westermann and Julia Schutz Veiga*

Comparison of Embedding Methods for Retrieval Under Noisy Institutional Labels 324
*Tereza Novotná and Jakub Harašta*

PONK: Tool for Client-Oriented Legal Writing in Czech 330
*Tereza Novotná, Jan Černý, Ivan Kraus, Ivana Kvapilíková, Jiří Mírovský, Arnold Stanovský and Barbora Hladká*

EmoLawConnect: An AI-Powered Matching Engine for Professionals Supporting Family Conflict Resolution 337
*Kohei Oshio*

Defending the Hierarchical Result Models of Precedential Constraint 344
*Henry Prakken and Wijnand van Woerkom*

Anomaly Detection in Legal Text – A Case Study on Cybersecurity Contracts 350
*Pavol Sokol, Laura Bachňáková Rózenfeldová, Dávid Varga, Simona Rudohradská and Regina Hučková*

An Expert-Validated LLM Framework for Transforming Legal Procurement Texts into Actionable Data 356
*Ivan Spada, Roberto Nai, Davide Audrito, Vittoria Sofia Margherita Trifiletti and Emilio Sulis*

Label Indeterminacy in AI & Law 364
*Cor Steging and Tadeusz Zbiegień*

From Court Decisions to Guiding Principles: Advancing Complex Legal Summarization with LLMs 371
*May Myo Zin, Ken Satoh and Georg Borges*

**Posters**

LR$^2$: A Legal RAG for Legal Reasoning over Cases 378
*Irene Benedetto, Luca Cagliero and Francesco Tarasconi*

Discovering the Potential of LLMs in Annotating Legal Texts for Argument Mining (Extended Abstract) 382
*Christina Berghegger, César Philippe, Karla Salas-Jimenez, Jean-Guy Mailly, Leila Moudjari and Laurent Perrussel*

Pro-Judice: Aligning LLMs with Procedural Fairness in Judicial Contexts 385
*Qingjing Chen, Rongxin Cheng, Yan Liu, Siyuan Zheng, Antonino Rotolo, Yun Liu and Weixing Shen*

Towards Interpretable Hierarchical Graph Neural Networks in Criminal Law 389
*Thiago Dal Pont, Isabela Sabo, Maite Vieira, Jomi Hübner and Aires Rover*

A Tool for Handling Unexpected Exceptions in Legal Reasoning 392
*Thomas Ecobichon, Mathieu Carpentier, Sylvie Doutre and Jean-Guy Mailly*

Normative Risk Assessment in the AI Act 395
*Beatrice Ferrigno, Marco Billi, Yasaman Yousefi and Antonino Rotolo*

An Argumentative Explanation Framework for Generalized Reason Model with Inconsistent Precedents 398
*Wachara Fungwacharakorn, Gauvain Bourgne and Ken Satoh*

Uncovering AI Governance Themes in EU Policies Using BERTopic and Thematic Analysis 400
*Delaram Golpayegani, Marta Lasek-Markey, Arjumand Younus, Aphra Kerr and Dave Lewis*

Foundations for Territorial Disambiguation in Law: A Preliminary Study Using the Education Act 2005 412
*Safia Kanwal, Livio Robaldo, Hafsa Dar, Davide Liga and Joseph Anim*

Extracting Unwritten Sentencing Rules from Czech Higher Courts' Verdicts 415
*Tomáš Knap*

LeAK-GraphRAG: A Legal Academic Knowledge GraphRAG Framework with Hybrid Retrieval for Mitigating Hallucinations 418
*Jia Liu and Jianhao Tang*

A GraphRAG Approach to Enhancing the Ability of Causal Reasoning in Autonomous Driving Cases 421
*Jia Liu and Xingtong Chen*

Automated Identification of Criminal Offences from Case Descriptions Using Ontologies, Rules and LLMs 424
*Jelena Matković, Marko Marković and Stevan Gostojić*

Principles of Law: LLMs vs RegEx 427
*Marianna Molinari, Ilaria Angela Amantea, Marinella Quaranta and Guido Governatori*

Packaging Thematic Analysis as an AI Workflow for Legal Research 430
*Vojtěch Pour*

Plans and Diversions EA 433
*Galileo Sartor, Guido Governatori, Giuseppe Pisano, Antonino Rotolo and Adam Wyner*

Enhancing Automated Decision-Making in Administrative Law Through Digital Twins of Legislation. A Grounded Theory Approach 436
*Florian Schnitzhofer and Christoph G. Schuetz*

Modeling U.S. Supreme Court Briefs with Computational Argumentation 439
*Heng Zheng, Dexter Williams and Bertram Ludäscher*

Author Index 442

# Long Papers

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251572*

# Precedent as Interpretive Resource: A Case Frame Analysis of Statutory Reasoning

Michał Araszkiewicz[a,1]
[a]*Jagiellonian University in Kraków, Department of Legal Theory*
ORCiD ID: 0000-0003-2524-3976

**Abstract.** Research in AI and Law has traditionally focused on case-based reasoning in common law systems, where factual similarity and factor-based modeling dominate. Much less attention has been given to continental jurisdictions, where judicial decisions are formally non-binding but still exert significant interpretive influence. This paper investigates how precedents function as interpretive resources in statutory reasoning, using the Polish Supreme Court decision as a case study. We use the case frame method, which structures judicial reasoning into slots: interpreted provision, competing interpretations, supporting arguments, meta-rules, and final decision. By applying this model to the source case and its reception in later case law, we identify three recurring patterns: (i) cross-domain influence, where the interpretive meta-rule travels across criminal, administrative, and transitional justice contexts; (ii) different takeaways, where distinct interpretive holdings (e.g., evidentiary standards, procedural guarantees) are extracted from the same judgment; and (iii) modificatory application, where courts cite the precedent but reshape its operative force.

**Keywords.** Case-based reasoning, Case frames, Continental legal systems, Precedent, Statutory interpretation

## 1. Introduction

AI and Law research has developed rich models of case-based reasoning (CBR) in the context of common law jurisdictions, where stare decisis makes earlier judicial decisions formally binding. Much of this work has focused on factor-based representations of cases and on methods for reasoning about similarities and differences between fact situations [1-8]. By contrast, continental legal systems, where statutes are primary sources of law and precedents are formally non-binding, have received much less attention in case-based modeling until recently [9, 10]. Here, the use of earlier cases arises not primarily from factual analogy but from interpretive reasoning: courts cite and rely on prior decisions as persuasive authorities for construing statutory terms. Existing factor-based CBR models are ill-suited to capture this practice, since the decisive question is typically not whether current facts resemble those in earlier cases, but how statutory provisions were interpreted and eventually applied.

To model such reasoning, this paper adopts the Case Frames approach [10], which represents cases in terms of their interpretive structure: the provision at stake, the interpretive methods employed, the competing readings considered, the meta-rules invoked, and the final decision. Case frames allow us to see how courts make use of earlier judgments in statutory interpretation.

[1] Corresponding Author: michal.araszkiewicz@uj.edu.pl. This research was funded by National Science Centre, Poland (grant no. 2023/49/B/HS5/00753).

The contribution of this paper is as follows. It demonstrates the effectiveness of the case frames method in structurally presenting the content of cases under continental law and in tracing how subsequent courts make use of earlier decisions. Our manually crafted dataset, consisting of 33 judgments citing the landmark Polish Supreme Court decision V KK 74/08,[2] reveals three recurring patterns: (i) cross-domain transfer of an interpretive meta-rule, (ii) reliance on different holdings embedded in the same decision, and (iii) modification of the transferred meta-rule. Together, these patterns highlight argumentation practices in continental legal culture that differ significantly from those in common law and contribute to the theoretical foundations of AI and Law by enabling the study of distinctive forms of precedent use in statutory systems.

The paper is structured as follows. Section 2 introduces the case frame method and explains the simplified representation employed in our analysis. Section 3 presents the source case *V KK 74/08* and reconstructs its interpretive content. Section 4 examines its reception in later judgments, distinguishing three characteristic modes of citation and adaptation. Section 5 discusses the implications for AI and Law, situating our findings within the literature on case-based and rule-based reasoning. Section 6 concludes and outlines directions for future work, including the integration of case frames with formal methods of legal knowledge representation including the Semantic Web technologies, and large language models.

## 2. Case Frames Method

In statutory legal systems, once the evidentiary facts are established, the crucial legal question usually turns on how to interpret the wording of the applicable statute. Precedents therefore function less as repositories of factual patterns and more as sources of interpretive knowledge. To capture this knowledge, we use the case frame method, a structured representation of judicial reasoning in interpretive disputes.

For the purposes of this study, we employ a simplified version of the case frame model [10]. Each case frame records the following elements:

- Case name: unique code identifying the case.
- Interpreted provision: the statutory clause or term at issue.
- Competing interpretations: alternative readings considered by the court.
- Arguments: the interpretive methods invoked (e.g., linguistic, systemic, purposive).
- Meta-rule: the second-order directive governing the relationship between methods (e.g., primacy of linguistic interpretation, holistic combination).
- Final decision: the interpretation ultimately adopted.

This representation allows us to observe how courts cite and reuse earlier cases. A later decision may transfer a case frame in its entirety, but often only selected elements are borrowed. For example, a court might cite a precedent to adopt its meta-rule while ignoring its specific interpretation of a provision, or conversely, invoke the same statutory interpretation but change the justificatory arguments. Case frames thus provide a fine-grained lens through which to study cross-domain transfers and modifications of precedents in continental legal jurisdictions.

[2] There exist cases in Polish law that are quoted much more frequently. This case was selected randomly from the pool of cases that have more than 20 citations.

In the sections that follow, we apply this model to analyze the reception of a landmark Polish Supreme Court's judgment. By reconstructing case frames for the source case and for the later cases that cite it, we can show how interpretive knowledge travels across doctrinal boundaries, sometimes intact, sometimes transformed.

## 3. Source Case: V KK 74/08

The judgment V KK 74/08 of the Polish Supreme Court (18 November 2008) arose from a cassation in a homicide case. The defendant had been convicted of murder, and one of the key issues concerned the recidivist aggravator in art. 148 § 3 k.k. This provision imposes harsher penalties when the perpetrator commits homicide after having already been convicted of "murder" (PL: zabójstwo). The problem was whether a prior conviction for a privileged form of homicide — for example, killing under strong emotional disturbance — also counts as "murder" for the purpose of this aggravator.

This was not a trivial question. If "zabójstwo" were read broadly to cover all forms of homicide in the Code, then even someone previously convicted in mitigated circumstances would face aggravated punishment if convicted again. But if the term were read narrowly, limiting it to the basic and qualified types of murder, then those with a privileged prior conviction would not fall under the aggravator. The interpretive dispute thus turned on whether the legislature intended "zabójstwo" to be a blanket reference to all types of homicide, or a more selective reference consistent with the internal structure of the Criminal Code.

The Supreme Court's resolution of this issue made V KK 74/08 a landmark decision. It began with a linguistic analysis, noting that the wording could be understood in more than one way. Because the linguistic outcome was not univocal, the Court turned to systemic reasoning: privileged homicide provisions exist precisely to mitigate liability, and treating them as aggravating factors would contradict their function. From this, the Court drew the broader principle that linguistic interpretation is the starting point but not the end point of statutory reasoning. Even where terms appear clear, their meaning must be tested against systemic and purposive considerations. Let us note that this is one among many possible meta-rules governing the sequence of application and the preference relations between different interpretation canons [11,12].

Beyond this interpretive canon, the judgment also addressed procedural guarantees (prohibition of *reformatio in peius*), the standard of proof in circumstantial cases, and the scope of clerical error correction. These holdings, although ancillary in the original case, would later be invoked independently by other courts. The following Table expresses the crucial interpretive problem of the case.

| Slot | Content |
|---|---|
| **Case name** | Case: V KK 74/08, Supreme Court (Poland), 19 November 2008 |
| **Interpreted provision** | Art. 148 §3 Criminal Code: *"If the perpetrator of homicide was previously convicted of homicide, …"* |
| **Competing interpretations** | Broad interpretation: "homicide" covers all forms under Art. 148, including privileged types (e.g. infanticide, euthanasia). |

| Slot | Content |
|---|---|
| **Arguments** | Narrow interpretation: "homicide" covers only basic and aggravated forms; privileged types are excluded.<br>Linguistic: the wording "homicide" is ambiguous, could support both broad and narrow readings.<br>Systemic: excluding privileged forms preserves coherence of the Criminal Code.<br>Purposive: privileged forms are meant to mitigate liability, inconsistent with aggravation. |
| **Meta-rule** | Holistic canon: linguistic interpretation has priority but not exclusivity; even clear terms must be tested against systemic and purposive considerations. |
| **Main holding** | Narrow interpretation adopted: Art. 148 §3 applies only to prior convictions for basic and aggravated homicide; privileged forms do not trigger the aggravating provision. |
| **Additional holdings** | Proof standard: conviction on circumstantial evidence valid only if all reasonable alternatives are excluded.<br>Guarantee: prohibition of *reformatio in peius* reaffirmed.<br>Clerical errors: art. 105 Code of Criminal Procedure allows correction only of genuine clerical mistakes, not substantive changes. |

Table 1. Case Frame representing the Source Case.

## 4. Reception of V KK 74/08 in Later Case Law

The significance of V KK 74/08 lies not only in its original resolution of a homicide provision, but also in the way later courts have cited and adapted it. In Continental legal systems, precedents are formally non-binding, yet they often serve as interpretive resources. What is striking in this line of reception is not merely that later courts quote V KK 74/08, but that they do so in markedly different ways. In order to trace the reception of V KK 74/08, we created a dataset of 33 cases citing it. The citing cases were identified on the basis of hyperlink system available in the LEX database: one of the leading electronic commercial repositories of Polish legal resources. All citations of the source cases were identified and their structure analyzed by an experienced lawyer. Three patterns emerge:

**Cross-domain influence** – the case is cited outside criminal law, for example in administrative and transitional justice proceedings, illustrating that precedents in continental legal culture can be mobilized across doctrinal boundaries.

**Different takeaways from the same case** – later courts may extract distinct elements of the case frame (e.g., the interpretive meta-rule, the standard of proof, or the guarantee against *reformatio in peius*), treating them as independent resources.

**Modificatory application** – courts sometimes narrow or transform the very element they cite, producing outcomes that differ in spirit from the original decision.

Each of these phenomena can be represented through case frames, which make visible which elements are transferred, which are omitted, and which are reshaped. In the following subsections, we illustrate each pattern with examples.

### *4.1. Cross-Domain Influence*

The most striking feature of the reception of V KK 74/08 is the extent to which its interpretive meta-rule, according to which linguistic primacy does not imply exclusivity, and that even seemingly clear terms must be tested against systemic and purposive arguments, has been mobilized beyond the original context of homicide law.

In criminal law, the decision has been cited not only in later homicide cases, but also in disputes concerning other provisions of the Criminal Code, such as the interpretation of "possession" in drug offences or of "significant quantity" in narcotics law. In each instance, courts explicitly referred to V KK 74/08 to justify the methodological move from literal wording to systemic and purposive arguments. This shows that the meta-rule has become a general meta-rule of statutory interpretation, detached from the original dispute over homicide provision.

The meta-rule has also travelled into administrative adjudication, where courts and other authorities have faced questions of statutory interpretation under regulatory statutes. In such settings, V KK 74/08 has been cited to stress that reliance on apparent linguistic clarity is insufficient: the statutory term must be construed in considering its systemic context and purpose.

The precedent has influenced cases concerning the status of repressed persons under the 2015 Act on opposition activists. Here, the term "harm" (*szkoda*) in art. 3 pkt 3 lit. b raised an analogous interpretive issue: does it require actual realized detriment, or can politically motivated attempts at repression also qualify? By invoking V KK 74/08, the Supreme Administrative Court underscored that interpretation could not stop with the linguistic reading but had to integrate systemic and purposive considerations.

The following table illustrates how the overall structure of the source case reasoning with the meta-rule was mirrored in two selected cases: one from the field of criminal law and another from the administrative law.

| Slot | II AKa 30/10 (Court of Appeal, Szczecin, 2010) | SKO 4134/2/11 (Self-Government Appeals Board, 2011) |
|---|---|---|
| **Case name** | Case: II AKa 30/10, Court of Appeal Szczecin, 2010 | Case: SKO 4134/2/11, Self-Government Appeals Board, 2011 |
| **Interpreted provision** | Drug law provision: prohibition of "possession of narcotics." | Environmental protection law: definition of "environmental nuisance" in relation to odor. |
| **Competing interpretations** | Narrow: "possession" = durable, significant possession only. Broad: includes immediate-use possession. | Narrow: odor not explicitly listed, so excluded. Broad: odor falls within the scope of environmental nuisance. |
| **Arguments** | Linguistic: "possession" ambiguous regarding minimal amounts. Systemic: excluding immediate-use creates | Linguistic: "odor" not literally covered. Systemic: environmental law protects against all nuisances affecting quality of life. Purposive: |

| Slot | II AKa 30/10 (Court of Appeal, Szczecin, 2010) | SKO 4134/2/11 (Self-Government Appeals Board, 2011) |
|---|---|---|
| | loopholes. Purposive: aim is to prevent all narcotic circulation. | statutory aim to protect human health and environment includes odors. |
| **Meta-rule** | Holistic canon (citing V KK 74/08): linguistic reading not decisive; systemic and purposive arguments decisive. | Holistic canon (citing V KK 74/08): literal reading not conclusive; systemic and purposive reasoning required. |
| **Final decision** | Broad interpretation: possession includes immediate-use quantities; law applies regardless of possession purpose. | Broad interpretation: odor included as an environmental nuisance; protection extends to air quality. |

Table 2. Exemplary transfer of the interpretive meta-rule to subsequent cases.

These two cases illustrate how the holistic canon of V KK 74/08 travels across doctrinal boundaries. In II AKa 30/10**,** the Court of Appeal used it to expand the meaning of "possession" in drug law, while in SKO 4134/2/11 an administrative body relied on it to include odor within environmental nuisances. Despite the differences in subject matter, both decisions demonstrate the same reasoning pattern: when linguistic interpretation leaves uncertainty or gaps, systemic coherence and legislative purpose drive the outcome.

The faithful transfer of the holistic canon from V KK 74/08 does not lock courts into a uniform interpretive direction. Once systemic and purposive arguments are weighed alongside the linguistic text, the result may be either a broadening or a narrowing of the statutory term. Both directions remain consistent with the meta-rule, as its role is not to dictate substance but to ensure methodological completeness. The contrast between I KZP 9/11 and III OSK 732/21 illustrates this dual potential.

| Slot | Broadening outcome | Narrowing outcome |
|---|---|---|
| **Case name** | I KZP 9/11 (Supreme Court, 2011) | III OSK 732/21 (Supreme Administrative Court, 2022) |
| **Interpreted provision** | Criminal law provisions on security measures – whether their duration is capped by the length of the imposed sentence | Art. 3(3)(b) Act of 2015 on opposition activists ("harm" caused by offence or petty offence) |
| **Competing interpretations** | Narrow: security measures may not exceed sentence duration. Broad: security measures can last independently of sentence duration. | Broad: "harm" includes potential or attempted detriment. Narrow: "harm" requires actual, realized detriment. |
| **Application of meta-rule** | Court applied holistic canon: systemic and purposive considerations (protection of | Court applied holistic canon: purposive aim of recognizing genuine repression reinforced |

| Slot | Broadening outcome | Narrowing outcome |
|---|---|---|
| | society, coherence of security law) outweighed the literal link to sentence duration. | a restrictive reading of "harm." |
| **Effect** | **Broadened interpretation**: security measures not capped by sentence length, extending their applicability. | **Narrowed interpretation**: only actual, realized harm qualifies, excluding potential or attempted repression. |

Table 3. Faithful transfer of interpretive meta-rule: broadening or narrowing outcomes.

In sum, the reception of V KK 74/08 across criminal, administrative, and transitional justice cases confirms both the portability and flexibility of its holistic canon. The same methodological directive—testing linguistic meaning against systemic and purposive considerations—travels easily across domains yet does not dictate uniform results. Depending on the balance of arguments, it may justify either a **broadening** (as in I KZP 9/11 on security measures) or a **narrowing** (as in III OSK 732/21 on the notion of harm) of statutory terms. This dual character demonstrates that the meta-rule is used as a framework for disciplined reasoning, one that accommodates divergent but methodologically consistent outcomes.

*4.2. Different Takeaways from the Same Source Case*

While V KK 74/08 is most frequently cited for its holistic canon of interpretation, later courts have also drawn on its other holdings, each of which resolved an interpretive issue under different statutory provisions. These include the standard of proof for circumstantial evidence (art. 7 CCP), the prohibition of *reformatio in peius* (art. 443 CCP), and the scope of clerical error correction (art. 105 CCP).

Under continental law assumptions, such holdings have no binding force: they are ancillary to the main ruling and are not meant to structure future case law. Nevertheless, practice shows that they are transferred virtually 1:1 into subsequent decisions. Courts cite these passages of V KK 74/08 as if they were stable interpretive authorities, confirming their persuasive weight in shaping juridical consensus. The below table illustrates the three streams of the cases transferring the selected elements. Notably, they are all from the field of procedural criminal law. For the sake of conciseness, we do not present full case frames of the citing cases.

| Interpreted provision | Interpretation in V KK 74/08 | Later use in case law |
|---|---|---|
| **Art. 7 CCP (proof standard)** | Conviction on circumstantial evidence valid only if all reasonable alternatives are excluded. | Cited as the authoritative evidentiary standard in circumstantial cases. |
| **Art. 443 CCP (*reformatio in peius*)** | An appeal cannot worsen the defendant's position when only they appealed. | Reaffirmed in appellate practice as a core procedural guarantee. |

| Interpreted provision | Interpretation in V KK 74/08 | Later use in case law |
|---|---|---|
| **Art. 105 CCP (clerical error correction)** | Corrections limited to genuine clerical mistakes, not substantive changes. | Cited in procedural rulings on scope of permissible corrections. |

Table 4. Three streams of specific elements transfer in criminal case law quoting V KK 74/08.

*4.3. Modificatory application*

The mere fact that a court cites V KK 74/08 does not guarantee that the holistic interpretive canon is transferred in its original form. Citation can be a misleading signal: while the authority of the precedent is invoked, the substantive element being transferred may be reduced or redirected. A striking example comes from II AKa 447/12, where the appellate court formally acknowledged V KK 74/08 but confined the use of systemic and purposive interpretation to situations where linguistic interpretation itself does not resolve the interpretative issue—that is, only when the text itself allows at least two different readings. This move effectively narrowed the meta-rule, stripping it of its broader function in V KK 74/08, where systemic and purposive considerations were required even when the wording seemed clear. To present this phenomenon in a clear manner in the table below we contrast the narrowing reading of the meta-principle (as applied in II AKa 447/12) with faithful transfer of the meta-rule which took place in another case, V KK 119/10.

| Slot | II AKa 447/12 (Court of Appeal, Gdańsk, 2012) | V KK 119/10 (Supreme Court, 2010) |
|---|---|---|
| **Interpreted provision** | Art. 77 §2 CC – whether parole restrictions must be repeated in a combined sentence | Provisions on anonymous witnesses – whether "personal data" includes contextual identifiers |
| **Original canon from V KK 74/08** | Linguistic meaning has priority but not exclusivity; systemic and purposive reasoning must also be applied | Same holistic canon: linguistic reading tested against systemic and purposive arguments |
| **Application of canon** | Court cites *V KK 74/08* but restricts systemic and purposive reasoning to cases of explicit ambiguity in the text | Court cites *V KK 74/08* and fully applies the canon to broaden "personal data" beyond literal wording |
| **Final decision** | Meta-rule narrowed: linguistic interpretation regains near exclusivity unless two readings are possible | Meta-rule faithfully applied: systemic and purposive reasoning decisive in extending protection scope |

Table 5. Modifying vs. faithful transfer of the meta-interpretive rule.

This comparison illustrates that citation of V KK 74/08 can serve different strategic purposes. In some instances, as in II AKa 447/12, the precedent is invoked mainly to anchor authority, while its operative force is curtailed. In others, like V KK 119/10, the canon is transferred intact and used to expand the statutory meaning. Importantly, modification need not always be restrictive: courts could also make the canon more open-ended by broadening the range of permissible systemic or purposive arguments, or by introducing additional interpretive elements. What remains constant is the dual function of citation—both as a gesture of continuity with established authority and as a vehicle for recalibrating interpretive freedom.

## 5. Discussion

For AI and Law, our findings suggest the need to move beyond factor-based models of case-based reasoning that dominate analyses of common law precedents. Unlike the Anglo-American context, where the central task is identifying factual similarities and differences, in continental legal culture statutory law, the key unit of transfer is the interpretive element: a reading of a provision, a set of competing arguments and a meta-rule about how interpretation should proceed. Case frames provide a natural representational tool for this setting, as they decompose judgments into slots [13] that can be selectively reused in later cases.

This insight resonates with earlier work by Rissland and Skalak, who emphasized the importance of modeling open-textured concepts and flexible interpretive strategies in case-based reasoning systems (CABARET) [14, 15]. Their approach underscored that legal reasoning frequently involves applying or modifying abstract interpretive heuristics rather than simply matching factual patterns. Our analysis confirms this observation within a statutory, continental legal culture: courts do not merely "follow facts," but rather transfer, reframe, and sometimes narrow or broaden interpretive statements and meta-rules across doctrinal boundaries.

The semi-formal representation developed here complements these insights. By explicitly representing interpretive slots—provisions, competing interpretations, arguments, meta-rules, holdings—case frames provide a richer structure for encoding how open-textured legal terms are resolved in practice. Recent research highlights that large language models (LLMs) can be leveraged to identify and reason about open-textured terms in statutes [16]. Our approach complements this line of work: whereas LLMs can help detect where open-texture arises, case frames capture how courts *resolve* such open-texture by mobilizing interpretive arguments and meta-rules. Such research should also relate to the earlier work concerning identification of elements of case law useful for statutory interpretation [17].

A further point concerns methodology. All the tables presented in this paper except for the first one were in fact generated with the assistance of ChatGPT-5 from the raw text of the judgments, using carefully designed prompts explaining the Case Frames method,[3] and subsequently verified by an expert lawyer. This workflow demonstrates the feasibility of combining LLM-supported extraction with expert supervision in building structured representations of case law. We intentionally refrain from proposing

[3] The key prompts are available here: https://docs.google.com/document/d/1Zd_V8mYkiWEkhydFBORI4TQ48mqphMElQm7POB3QKdQ/edit?usp=sharing

a fully formal model here, as our dataset is based on raw textual documents. Nevertheless, existing formal frameworks of legal reasoning [18 - 21] could be directly tested against, or integrated with, our semi-formal structures.

## 6. Conclusions and Further Research

This paper has shown how the case frames method can capture the structural reuse of interpretive reasoning in continental legal systems, using the landmark decision *V KK 74/08* and 33 judgments citing it as a dataset. The analysis identified three recurring patterns of reception—cross-domain transfer of interpretive canons, reliance on additional holdings, and modificatory applications—which together demonstrate that precedents in statutory cultures function primarily as sources of interpretive resources rather than binding commands. Case frames, by decomposing cases into provision, competing interpretations, arguments, meta-rules, and holdings, offer a transparent, systematic and replicable way to study these dynamics.

A further contribution of this work lies in its methodological direction, which appears promising for practical applications. The case frame tables were generated with the assistance of ChatGPT-5, based on carefully designed prompts explaining the method, and then validated by an expert lawyer. This workflow illustrates how LLMs can assist in extracting structured representations from raw judicial texts, while human expertise ensures legal accuracy. The result is a semi-formal dataset that bridges the gap between textual judgments and formal models of legal reasoning, motivating the development of hybrid legal applications useful for legal practice [22].

Although primarily analytical rather than experimental, this research provides a foundation for systematic experimentation with LLMs on case datasets. Future work could explore how models can be trained or fine-tuned to generate reliable case frames automatically, to detect interpretive issues, and to predict which elements of a source case are likely to be transferred, modified, or ignored in later decisions. This would complement ongoing work on open-textured statutory terms by enabling computational tools not only to identify interpretive uncertainty but also to model how courts resolve it.

Further research should also pursue three broader directions. First, building a larger corpus of case frames would allow for quantitative analysis of how interpretive arguments and meta-rules propagate across domains. Second, aligning case frames with formal knowledge representation methods and reasoning models would make it possible to test the expressive power of existing solutions (e.g. [23]) against real-world interpretive practices. Third, by encoding case frames in machine-readable formats (e.g. RDF), they could be integrated into Semantic Web environments and leveraged in retrieval-augmented generation systems [24], offering structured, high-quality resources for AI systems trained on legal texts.

Taken together, these avenues point toward a hybrid research agenda: case frames provide a jurisprudentially meaningful unit of analysis while also serving as a computationally useful resource, bridging doctrinal insight and AI-supported modeling of statutory interpretation in continental legal systems.

## References


[1] Ashley K. *Modeling legal arguments: reasoning with cases and hypotheticals.* MIT Press; 1991.

[2] Bench-Capon T. HYPO's legacy: introduction to the virtual special issue. *Artif Intell Law.* 2017; 25:205-250.

[3] Horty JF. Reasons and precedent. In: *Proceedings of the 13th International Conference on Artificial Intelligence and Law – ICAIL 2011.* p. 41-50.

[4] Horty JF, Bench-Capon T. A factor-based definition of precedential constraint. *Artif Intell Law.* 2012; 20(2):181-214.

[5] Rigoni A. An improved factor-based approach to precedential constraint. *Artif Intell Law.* 2015; 23:133-160.

[6] Prakken F. A formal analysis of some factor- and precedent-based accounts of precedential constraint. *Artif Intell Law.*2021; 29(4):559-585.

[7] Aleven V. *Teaching case-based argumentation through a model and examples.* PhD thesis, University of Pittsburgh; 1997.

[8] Bench-Capon T, Atkinson K. Precedential constraint: the role of issues. In: *Proceedings of the 18th International Conference on Artificial Intelligence and Law – ICAIL 2021.* p. 12-21.

[9] Araszkiewicz M. A hybrid model of argument concerning preferences between statutory interpretation canons. In: Francesconi E, Borges G, Sorge C, editors. *Legal knowledge and information systems – JURIX 2022: The Thirty-fifth Annual Conference.* IOS Press; 2022. p. 3-12.

[10] Araszkiewicz M. Case Frames and case-based arguments in statutory interpretation. In: Savelka J, Harasta J, Novotna T, Misek J, editors. *Legal knowledge and information systems – JURIX 2024: The Thirty-seventh Annual Conference.* IOS Press; 2024. p. 2-13.

[11] Walton D, Sartor G, Macagno F. An argumentation framework for contested cases of statutory interpretation. *Artif Intell Law.* 2016; 24:51-59.

[12] MacCormick N, Summers R, editors. *Interpreting statutes.* Routledge; 1991.

[13] Minsky M. A framework for representing knowledge. In: Winston PH, editor. *Psychology of computer vision.* MIT Press; 1975. p. 211-277.

[14] Skalak DB, Rissland EL. Arguments and cases: an inevitable intertwining. *Artif Intell Law.* 1992; 1(1):3-44.

[15] Rissland EL, Skalak DB. CABARET: rule interpretation in a hybrid architecture. *Int J Man-Mach Stud.* 1991; 34(6):839-887.

[16] Guitton C, Tamò-Larrieux A, Gubelmann R, Karray G, Mayer S. Identifying open-texture in regulations using LLMs. *Artif Intell Law.* 2025. https://doi.org/10.1007/s10506-025-09450-0

[17] Šavelka J, Ashley KD. Legal information retrieval for understanding statutory terms. *Artif Intell Law.* 2022; 30(2):245-289.

[18] Prakken H. An abstract framework for argumentation with structured arguments. *Argument Comput.* 2010; 1(2):93-124.

[19] Gordon TF, Prakken P, Walton D. The Carneades model of argument and burden of proof. *Artif Intell.* 2007; 171(10-11):875-896.

[20] Prakken H, Sartor G. A formal framework for combining legal reasoning methods. In: *Proceedings of the 19th International Conference on Artificial Intelligence and Law – ICAIL 2023.* p. 227-236.

[21] Atkinson K, Bench-Capon T. ANGELIC II: an improved methodology for representing legal domain knowledge. In: *Proceedings of the 19th International Conference on Artificial Intelligence and Law – ICAIL 2023.* p. 12-21.
[22] Ashley K. *Artificial intelligence and legal analytics: new tools for law practice in the digital age.* Cambridge University Press; 2017.
[23] Zin MM, Nguyen HT, Satoh K, Sugawara S, Nishino F. Improving translation of case descriptions into logical fact formulas using LegalCaseNER. In: *Proceedings of the 19th International Conference on Artificial Intelligence and Law – ICAIL 2023.* ACM Press; 2023. p. 462-466. https://doi.org/10.1145/3594536.359514
[24] Francesconi E. A description logic framework for advanced accessing and reasoning over normative provisions. *Artif Intell Law.* 2014; 22(3):291-311.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251573*

# JuriDiscovery: Automated Analysis of Legal Dispute Triggered by Public Policies Using Artificial Intelligence

*A Case Study Applied to the Individual Microentrepreneur*

Loreane BRANDIZZI [a,1] Guilherme RODRIGUES [a,2]
[a] *University of Brasília, Brazil*

**Abstract.** The litigation arising from weaknesses in the implementation of public policies in Brazil has become a growing challenge. Among the many affected groups, Individual Microentrepreneurs (MEI) represent a significant case, with thousands of lawsuits addressing rights, obligations, and structural shortcomings of this regime. Analyzing such cases is complex due to the large volume and unstructured nature of judicial data. Advances in Artificial Intelligence (AI), particularly Large Language Models (LLMs) and semantic clustering, offer opportunities to transform legal texts into structured knowledge. This study presents a pipeline that combines BERTopic and prompt engineering to extract and analyze judicial cases involving MEIs. The results highlight recurring issues such as employment relationships, social security benefits, free legal aid, tax exemptions, and registration fraud. The methodology proved robust, scalable, and adaptable to other judicial contexts, contributing to both academic research and decision-making in public administration and legal practice.

**Keywords.** Jurimetrics, Legal Analytics, Public Policies, Natural Language Processing , Large Language Models, Topic Modeling

## 1. Introduction

Guided by its Digital Government Strategy, Brazil has advanced a transformative agenda for public administration and the economy, ranking 14th worldwide in the United Nations 2023 index of digital public services [1]. Yet, evaluating Brazilian public policies remains complex: traditional methods such as surveys, while robust, are costly, time-consuming, and unable to capture the dynamic nature of social impacts [2]. This gap makes policy evaluation itself a structural challenge. The Individual Microentrepreneur regime, established by Complementary Law No. 128/2008 [3], is an example of a complex public policy. It allows self-employed workers and small business owners to formalize their activities under a simplified tax and social security framework. Fifteen years after its creation, the regime has generated a lasting legacy of judicial disputes, making it a relevant case study for this research.

[1] e-mail: loreane.brandizzi@gmail.com
[2] e-mail: guilhermerodrigues@unb.br

Rather than relying solely on traditional surveys and economic models, policy evaluation can benefit from judicial data, where conflicts and system failures are formally recorded. Legal disputes often provide an X-ray of unresolved problems, as illustrated by the billion-dollar fraud in unlawful social security discounts at Brazil's INSS, which had been signaled for years through thousands of lawsuits and complaints before erupting in April 2025 [4,5,6] . To detect hidden problems of this scale before they erupt, jurimetrics, combined with advances in AI, offers transformative potential. The integration of statistical methods with NLP and LLMs has revolutionized legal data analysis [7,8], and this work contributes by applying LLMs to unstructured judicial data and enhancing the semantic representation of case summaries.

This work proposes to analyze the impact of the MEI policy within the judiciary and to show how modern AI techniques can automatically uncover the main issues arising from its implementation. Our main contribution is the development of a pipeline that applies Topic Extraction, enhanced with LLM-based prompt techniques, to transform thousands of unstructured judicial decisions into semantically organized and interpretable information. The method improves topic modeling by generating clearer and less redundant labels, enabling the identification of litigation patterns linked to public policy implementation. In addition to the analytical pipeline, we also provide a visualization tool that makes the results accessible to stakeholders interested in evaluating public policies.

The proposed approach is parameterizable, enabling its adaptation to different contexts of policy judicialization while supporting evidence-based decision-making.

## 2. Related Work

Jurimetrics combines statistical methods and data science to Law [9]. In Brazil, studies highlight its use in legal research [10] and in decision support [11]. More recently, the field has advanced with the application of AI and LLMs [7,8,12,13].

The rise of LLMs has accelerated the use of AI in jurimetrics, enabled by the Transformer architecture [14], advances in computational infrastructure, and the availability of large-scale datasets. In the legal domain, specific models such as LawGPT, ChatLaw, LexiLaw, and JurisLMs have been developed for tasks like summarization and decision classification [7,8]. In Portuguese, initiatives such as LegalBERT-pt [12] and JurisBERT [13] show the adaptation of these models to local contexts. Recent works, including CaseEncoder [15], SAILER [16], and LawGPT [17], highlight the growing sophistication and diversification of legal-domain language model architectures.

Despite progress, unstructured legal data remain difficult to process due to their lack of standardization and inherent complexity [7]. This challenge, recognized by the Brazilian Judiciary in the recent Manual for Standardization of Case Summaries [18], underscores the need for specialized modeling and data curation.

LLMs are increasingly used in legal data pipelines for tasks such as summarization and information extraction, supported by prompt-engineering strategies like Prompt Chaining [19], Chain-of-Thought [20], and In-Context Learning [21], enabling zero-shot summarization without explicit training [22,23].

The use of LLMs in jurimetrics is well established; however, their exclusive adoption raises concerns about interpretability, bias and reproducibility [24]. To mitigate these issues, unsupervised topic modeling remains a alternative for large-scale legal text anal-

ysis. Among such methods, BERTopic [25] stands out for generating semantically coherent and interpretable clusters, with comparative studies confirming its superiority over classical models such as LDA, NMF, and Top2Vec [26]. BERTopic can be enhanced through the integration of LLMs, which refine topic labeling via prompt-based methods that produce clearer labels [25]. This same synergy has been explored in recent works such as TopicGPT [27] and TNT-LLM [28].

Recent studies confirm BERTopic's effectiveness in legal text analysis across jurisdictions. Didwania et al. [29] compared it with LDA and NMF on Indian and UK judicial documents, finding similar coherence scores but higher semantic quality in expert evaluations. Likewise, Aouichaty et al. [30] analyzed nearly 200,000 Arabic legal judgments, showing that domain-specific monolingual embeddings achieved higher topic coherence and, overall, confirming BERTopic's suitability for large-scale legal corpora even when using standard Uniform Manifold Approximation and Projection (UMAP) and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) settings.

Building upon these insights, the present study adopts BERTopic for its maturity and interpretability, combining it with domain-adapted prompts for label refinement, summarization, and structured insight generation in Portuguese judicial texts, using **BERTimbau** as the embedding model [31].

## 3. Methodology

The proposed methodology is parameterized, allowing easy adaptation to other datasets. It consists of five stages (Figure 1) and is based on 57,075 Brazilian judicial case summaries explicitly mentioning "Microempreendedor Individual". The data were web-scraped from public sources and stored in a database, containing only case summaries and metadata (case number and court), with no personally identifiable information.

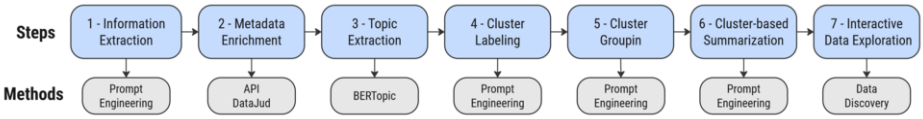



**Figure 1.** Executed stages of the pipeline.

Approximately 96% of the case summaries contained between 100 and 20,000 tokens, ensuring sufficient textual depth for analysis. Four large-context LLMs (`codestral-22b`, `pixtral-12b`, `deepseek-r1-distill-qwen-14b`, and `llama-3.1-8B-instruct`) were tested via the SERPRO [32] LLM API on a stratified sample of 200 cases using a common summarization prompt. Evaluations considered **format** and **content** criteria, with the latter verified through an LLM-as-a-judge approach using `GPT-3.5`. `Pixtral-12B` [33] achieved the best balance and was selected for subsequent stages.

### *3.1. Information Extraction*

Exploratory analysis showed that many mentions of MEIs in the case summaries were contextual, appearing in legal citations or references to related provisions, rather than

indicating that the MEI was directly involved as a party. This required an initial filtering step to identify cases where the MEI effectively participated in the dispute.

To ensure consistency and extract relevant information from the summaries, the texts were standardized following the Manual for Standardization of Case Summaries [18]. A prompt, which included a one-shot example to guide formatting, was applied to both reformat the summaries and verify MEI involvement (Figure 2), confirming participation in **11,775** of the **57,075** analyzed cases.

**Prompt Used for Structured Legal Summarization**

INSTRUCTIONS: You act as a legal assistant with expertise in making complex judicial case summaries more understandable, using clear and accessible language. Your approach follows the guidelines established by the *National Pact for Plain Language in the Judiciary*, an initiative of the *National Council of Justice (CNJ)*. Your task is to transform the provided summary into a clear, direct, and easy-to-understand response. Do not introduce new information, arguments, legal reasoning, or conclusions that are not explicitly contained in the original text, and do not make speculative inferences. Respond exclusively in Portuguese and strictly follow the indicated format.

Based on the provided case summary (ementa), extract the following fields and return the answer in JSON format:

- "descricao_caso": Summarize what happened in the case, explaining the main facts and what was requested. Do not mention names of people or companies. Up to 300 words.
- "questoes_em_discussao": List the main points discussed in the judgment. Up to 300 words.
- "solucoes_propostas": Explain briefly how the issue was analyzed or resolved. Up to 200 words.
- "decisao": State the outcome of the case. Use only one of the following options: **"Accepted"**, **"Denied"**, **"Partially Accepted"**, or **"Under Review"**.
- "tese": If applicable, describe the main idea of the decision, such as a rule that may apply to similar cases. Up to 100 words.
- "envolve_mei": Indicate **"Yes"** or **"No"** depending on whether the case explicitly involves an *Individual Microentrepreneur (MEI)* as plaintiff or defendant. If it is only a generic or indirect reference, answer **"No"**.
- "mei_do_processo": If envolve_mei is **"Yes"**, state whether the MEI is the **"Plaintiff"** or the **"Defendant"**. If unclear, write **"None"**.

**Example of Expected Response:**

```
{"descricao_caso": "An Individual Microentrepreneur (MEI) filed a lawsuit against the municipality
  claiming undue tax charges not established by law.",
  "questoes_em_discussao": "The case discussed whether the charge had a legal basis and was
  proportional to the MEI's business activity.",
  "solucoes_propostas": "The court ruled the charge unlawful due to the absence of legal foundation
  and disproportionality to the activity.",
  "decisao": "Accepted",
  "tese": "A municipality cannot charge taxes from an MEI without legal basis and proportionality.",
  "envolve_mei": "Yes",
  "mei_do_processo": "Plaintiff"}
```

**Figure 2.** Mapping between Prompt Fields and the CNJ Ementa Model (translated from Portuguese) [18].

### *3.2. Metadata Enrichment*

Metadata were obtained through the public DataJud API [34], with endpoints identified according to each case's court of origin. For MEI-related cases, metadata were successfully retrieved for **9,786** records, including details such as case type, jurisdiction level, municipality code, judging body, legal subjects, filing date, and procedural history.

### *3.3. Topic Extraction*

The concatenated texts from the fields `descricao_caso` and `questoes_em_discussao` were vectorized using `BERTimbau` for all cases in which MEI involvement was confirmed. The resulting embeddings were then clustered semantically with the `BERTopic` library [25].

The BERTopic model was configured with its default hyperparameters, which combine dimensionality reduction through UMAP with the HDBSCAN algorithm. While HDBSCAN was used with its default settings, the UMAP parameters were customized (`n_neighbors=30`, `n_components=5`, `min_dist=0.15`, `metric='euclidean'`) to reduce noise and improve topic separation.

Following clustering, BERTopic identifies topics using a frequency-based TF-IDF representation restricted to each cluster. It can also leverage LLMs for topic labeling through prompts applied to centroid documents and keywords. In this study, that strategy was adopted with specific adaptations described in the next step.

### *3.4. Cluster Labeling:*

Cluster labeling was performed through prompt engineering, instructing the LLM to infer labels from a representative set of documents. The prompt integrated three components: (i) TF–IDF topic keywords, (ii) the texts of the documents closest to each cluster centroid, and (iii) concise instructions in Portuguese. We employed the `Pixtral-12B` model via an external API (outside the standard `BERTopic` pipeline), sending one request per cluster with the prompt illustrated in Fig. 3 (translated from Portuguese).

For each cluster, labeling was based on the documents nearest to its centroid (in terms of cosine distance). To ensure the representativeness of this subset, $k$ members were selected, defined as $k = \min\{\max(\lceil \rho n \rceil, 15), 100, n\}$, where $\rho = 0.30$ and $n$ is the cluster size.

**Prompt Used for Topic Labeling**

```
You are an intelligent topic extraction assistant specialized in naming topics in a short, clear, and user-friendly
way, based on representative texts and keywords.  Your goal is to create a concise label that captures the main
common issue discussed within the topic, making it easily identifiable by end users.

Follow the guidelines below:
- Do not include the term ''MEI'' in the topic name.
- The name must contain between 1 and 3 words.
- Do not mention names of people, places, or institutions.
### Sampled texts from this topic:
- [text_1] ....  [text_K]
### Keywords:  [keywords]

Return only the short name of the new topic, without explanations or additional punctuation.
```

**Figure 3.** Prompt for Topic Labeling (translated from Portuguese).

In an alternative to the standard `BERTopic` workflow, which uses another prompt and a limited set of integrated LLMs, we invoked `Pixtral-12B` externally, applying a custom prompt and contextualized examples for topic labeling. Representative documents were selected proportionally to cluster size and based on centroid proximity.

### *3.5. Cluster Grouping*

Once the topic clusters were obtained, a verification stage evaluated potential dimensionality reduction, as some topics presented semantically similar names. Embeddings of each topic label were generated using `BERTimbau`, and cosine similarities were computed to form an $N \times N$ matrix, where $N$ corresponds to the total number of topics. Topic pairs with similarity above **0.75** were then submitted to the `Pixtral-12B` model with a

legal prompt (Figure 4) to confirm whether they represented the same issue. The model produced binary responses ("1" or "0"), resulting in a final binary $N \times N$ matrix of thematic equivalence.

**Prompt Used for Topic Equivalence Verification**

```
You are a legal assistant. Decide whether the two topics below essentially address the same core issue,
disregarding wording variations. Respond with a single character only: ''1'' (same issue) or ''0'' (different
issues).
Topic A:
- Title: {topico_a}
- Short description: {desc_a}
Topic B:
- Title: {topico_b}
- Short description: {desc_b}
Respond only with ''1'' or ''0''.
```

**Figure 4.** Prompt for Identifying Whether Two Topics Represent the Same Legal Issue

After constructing the similarity and binary matrices, a graph-based grouping strategy was applied to merge topics representing the same legal issue. Each topic was treated as a node, and groups were incrementally formed through a greedy and strict process: a topic was included only if it was connected in the LLM validation matrix (`1`) and showed cosine similarity above **0.75** with all existing members. This procedure substantially reduced topic redundancy while preserving semantic precision. In each group, all related cases were updated in the database with the metadata of the most representative topic, defined as the one with the largest number of associated cases.

- **Group 1:** Jurisdiction of Small Claims Courts; Classification for Small Claims Courts; Legitimacy in Executions; Judicial Legitimacy; Legitimacy in Small Claims Court; Legitimacy in Special Courts.
- **Group 2:** Rural Retirement; Social Security Benefits; Contributions and Retirement.
- **Group 3:** Undue Charges; Fee Exemption for MEIs; Wrongful Credit Reporting.
- **Group 4:** Legal Aid; Free Justice; Free Legal Aid for Appeals.
- **Group 5:** Welfare Benefits; Benefits for People with Disabilities.
- **Group 6:** Classification of Insured Persons; Survivor's Pension.
- **Group 7:** Debt Collection; Debt and Conciliation.

### *3.6. Cluster Based Summarization*

To consolidate each cluster's main characteristics, a LLM-based summarization step was applied using the previously identified $k$ representative cases. Six descriptive elements were generated: a **label** (required after topic merging), a **short** and **long description**, the **main issues in dispute**, the **proposed solutions**, and the **underlying legal doctrines**.

Each element was produced through a dedicated query to the model, using the relevant fields (`descricao_caso`, `questoes_em_discussao`, `solucoes_propostas`, `tese` and TF-IDF `keywords`) as context combined with specific output instructions. The base prompt for this stage is shown in Figure 5.

The specific instructions varied according to the type of information requested, guiding the model to generate updated cluster labels , concise summaries, descriptive texts, lists of issues, proposed solutions, and legal doctrines, each with predefined length and format constraints.

**Custom Prompt for Legal Summarization**

```
You are a legal assistant with expertise in analyzing judicial decisions. Your role is to synthesize information
clearly, accessibly, and in an organized manner, so that anyone can quickly understand the key topics, issues, and
solutions discussed in a set of cases.

Based on the provided texts, generate the following information: [SPECIFIC INSTRUCTION]

### Sampled case texts:
- [text field from case 1]
...
- [text field from case k]

Respond in Portuguese, objectively, following exactly the requested format. Do not add any explanations beyond the
answer.
```

**Figure 5.** Custom Prompt for Legal Summarization

### *3.7. Interactive Data Exploration*

An interactive Tableau Public dashboard visualizes the main findings of the study. The single integrated view provides a comprehensive analysis of judicialization involving MEIs, leveraging metadata from DataJud. It enables exploration of case volumes, their distribution across judicial branches and municipalities, temporal evolution, and MEI participation. Users can also examine topics derived from clustering and summarization models, with access to short and long descriptions, key issues, proposed solutions, and predominant legal doctrines.

The interactive dashboard are publicly available at https://public.tableau.com/app/profile/loreane.brandizzi/viz/JuriDiscovery/VisoGeral and the full pipeline implementation at https://github.com/LoreBrandizzi/juridiscovery.

### *3.8. Ablation Study and Evaluation*

To assess the effectiveness of each methodological stage, an *ablation study* was conducted. The data were clustered with **BERTopic** while systematically varying three factors: dataset (full corpus vs. MEI-filtered cases), textual input (original *ementa* vs. summarized description and issues) and methodological techniques (without LLM labeling, with LLM labeling, and with LLM-based grouping).

Topic coherence across stages was measured using the Normalized Pointwise Mutual Information (NPMI) [35], which quantifies word co-occurrence within a fixed context window for a topic's top terms. Values range from $[-1, 1]$, with higher scores indicating stronger semantic associations and, therefore, more coherent topics.

## 4. Results

It was observed that **20.65%** of the available cases effectively involve MEIs, either as plaintiffs or defendants. Furthermore, the highest concentration of cases involving MEIs is found in the **State Courts** and **Labor Courts**, which together account for more than **82%** of the total, as shown in Table 1.

Among the **11,775** analyzed cases, **3,779** were classified as noise by the HDBSCAN algorithm, as they did not present enough semantic density to form consistent clusters.

**Table 1.** Distribution of cases by court type and MEI involvement

| Court Branch | Involving MEI | Not Involving MEI |
|---|---|---|
| Labor Court | 4,523 | 17,726 |
| State Court | 5,122 | 14,252 |
| Federal Court | 2,049 | 13,181 |
| Electoral Court | 81 | 141 |
| **Total** | **11,775** | **45,300** |

The remaining **7.996** cases were clustered by BERTopic into **57 distinct topics**, which, after label refinement via the LLM, were consolidated into **54 topics**. These were subsequently aggregated into a final set of **40 topics** that represent the main judicial disputes related to the MEI public policy. Notably, the top **12 topics** account for approximately **90%** of all classified cases.

Table 2 presents the results of the ablation study, showing that noise was progressively reduced while the number of topics decreased and NPMI values increased, indicating more coherent and consistent clusters.

**Table 2.** Ablation experiments with different methodological stages.

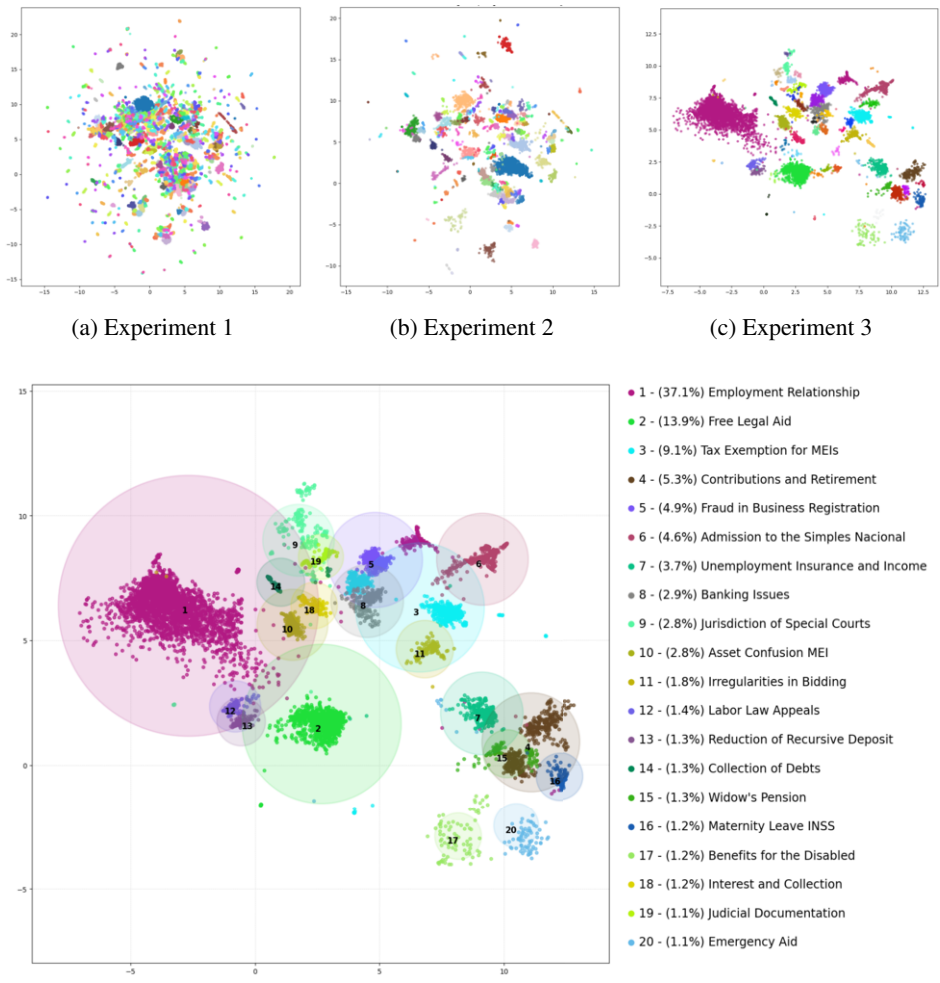
| Experiment | Records | Topics (no noise) | Noise | NPMI |
|---|---|---|---|---|
| **Experiment 1** | | | | |
| No filter (All processes) and Full summary | 53,256 | 523 | 22,065 | 0.16 |
| **Experiment 2** | | | | |
| Processes involving MEI and Full summary | 11,775 | 125 | 4,504 | 0.11 |
| **Experiment 3** | | | | |
| Processes involving MEI and Description + Issues in dispute | 11,775 | 57 | 3,779 | 0.32 |
| **Experiment 4** | | | | |
| Processes involving MEI, Description + Issues in dispute, and Cluster labeling + grouping | 11,775 | 40 | 3,779 | 0.43 |

Figure 6 illustrates the progressive refinement of topic clusters across Experiments 1–3, culminating in the final configuration of the 20 most representative topics.

The topic **Employment Relationship** clearly stands out as the dominant cluster, representing **37.1%** of the cases, followed by **Free Legal Aid** with **13,9%**, **Tax Exemption for MEIs** with **9.1%**, **Contributions and Retirement** with **5.3%** and **Fraud in Business Registration** with **4.9%**. These distributions reflect the 5 main areas of judicialization involving MEIs in Brazil.

The **Employment Relationship** topic concerns the legality of hiring individuals under MEI status, known as *pejotização*, a practice used to circumvent labor obligations [36]. Its relevance is underscored by the Brazilian Supreme Federal Court's nationwide suspension of related cases pending a ruling on Theme 1389 [37], and by the growing use of the term, which is even under consideration for inclusion in the Portuguese Language Orthographic Vocabulary [38].

Moreover, the **Free Legal Aid** topic involves disputes over MEIs' requests for court fee exemptions, which hinge on proving financial incapacity and on whether they are treated as individuals or legal entities.




(a) Experiment 1 (b) Experiment 2 (c) Experiment 3

(d) Experiment 4

**Figure 6.** The top panels illustrate the ablation experiments (1–4), showing the progressive organization of clusters. The bottom plot depicts the final configuration, where each point represents a judicial case projected in two dimensions from reduced embeddings of the summarized case ruling content. The circle size is proportional to the number of MEI-related cases classified under each respective topic (cluster).

Furthermore, the **Tax Exemption for MEIs** topic stems from regulatory reforms, especially Resolution No. 59/2020 of the Management Committee of Simples Nacional (CGSIM) [39], which exempted MEIs from municipal permits and licenses. The CGSIM is a governmental body responsible for deliberations and regulations affecting companies.

The **Contributions and Retirement** topic highlights the challenges faced in ensuring the effective realization of social security rights for MEIs. Although the MEI policy was designed to expand social protection and simplify social security contributions, the judicial cases reveal that many entrepreneurs encounter obstacles in proving contribution periods.

Finally, the **Fraud in Business Registration** topic concerns third parties fraudulently registering MEIs in the CNPJ without the owners' consent. The growing number of such cases prompted the Federal Revenue Service (RFB), in 2023, to launch a service enabling citizens to block unauthorized business registrations with their identification data [40].

## 5. Conclusion

### *5.1. Discussion*

The developed pipeline proved effective in generating automated and meaningful insights into the judicialization of public policies. By combining topic extraction techniques with LLMs, it transformed unstructured legal texts into interpretable analytical information. The approach performed well in the MEI policy context and, being parameterized, can be adapted to other judicial domains, supporting the identification and analysis of temporal, spatial and semantic patterns of litigation.

No manual validation or human labeling was used, ensuring reproducibility and adaptability to other contexts. Rather than revalidating `BERTopic`'s well-established performance, this study extends it to the legal domain, combining its clustering capabilities with LLM-based label refinement and semantic grouping.

The proposed pipeline progressively enhanced data quality and interpretability, with each stage [from prompt-based extraction to clustering, labeling, and grouping] incrementally refining the representation of judicial cases. A key methodological contribution lies in extending `BERTopic` through customized topic labeling and contextual grouping, which substantially improved clustering coherence and interpretability, as evidenced by the ablation study.

The integrated use of LLMs significantly enriched the labeling of topics and the generation of high-quality descriptive summaries, making the data more accessible, understandable and actionable for analysts, public managers and researchers.

### *5.2. Limitations and Future Work*

One limitation of this study concerns HDBSCAN's sensitivity, which classified 31% of cases as noise, reducing the analysis sample. Although prompt-based reclassification was tested, it compromised visualization clarity. Nevertheless, excluding borderline cases improved semantic homogeneity and coherence within clusters.

To mitigate LLM related risks (bias, misclassification, and over-simplification) four models were tested and evaluated by an independent LLM using the same prompt, ensuring the selection of the most accurate and contextually faithful model.

Another constraint stems from the dependence on the quality of case summaries. Very short entries were excluded to preserve analytical integrity. Despite being non-exhaustive, the dataset's size and diversity support reliable generalization of trends in MEI-related judicialization.

Future research may extend this pipeline to other public policy domains, such as social security, healthcare, and education.

## References


[1] D. S. Carvalho, P. A. de Castro, and M. Fernandes-Sobrinho, "Políticas públicas na era do governo digital: revisão de literatura," *GETEC*, 2024. In Portuguese.

[2] P. d. M. Jannuzzi, *Indicadores para diagnóstico, monitoramento e avaliação de programas sociais no Brasil*. IPEA, 2005.

[3] Brasil, "Lei complementar nº 128, de 19 de dezembro de 2008," 2008.

[4] F. de S. Paulo, "Aposentados do inss são vítimas de descontos indevidos há anos; veja a cronologia do caso," 2025.

[5] T. de Justiça do Distrito Federal e Territórios, "Aposentada deve ser indenizada por descontos indevidos em benefícios previdenciários," 2025.

[6] C. Jurídico, "Controle integrado e cidadão no foco: o caso dos descontos na folha dos aposentados," 2025.

[7] J. Lai, W. Gan, J. Wu, Z. Qi, and P. S. Yu, "Large language models in law: A survey," *AI Open*, vol. 5, pp. 181–196, 2024.

[8] G. Zambrano, "Case law as data : Prompt engineering strategies for case outcome extraction with large language models in a zero-shot setting," *Law, Technology and Humans*, vol. 6, no. 3, pp. 80–101, 2024.

[9] J. A. Z. Trecenti, "Diagramas de influência: uma aplicação em jurimetria," dissertação de mestrado, Universidade de São Paulo, Instituto de Matemática e Estatística, São Paulo, novembro 2015.

[10] M. Maia and C. A. Bezerra, "Análise bibliométrica dos artigos científicos de jurimetria publicados no brasil," *RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação*, vol. 18, pp. 1–23, 2020.

[11] E. E. S. Gonzales, "Trade dress: uma análise de jurimetria com ferramentas de inteligência artificial," dissertação de mestrado em propriedade intelectual e inovação, Instituto Nacional da Propriedade Industrial (INPI) – Academia de Propriedade Intelectual, Inovação e Desenvolvimento, Rio de Janeiro, Brasil, 2019.

[12] R. Silveira, C. Ponte, V. Almeida, V. Pinheiro, and V. Furtado, "Legalbert-pt: A pretrained language model for the brazilian portuguese legal domain," in *Anais da XII Brazilian Conference on Intelligent Systems*, (Porto Alegre, RS, Brasil), pp. 268–282, SBC, 2023.

[13] C. F. O. Viegas, B. C. Costa, and R. P. Ishii, "Jurisbert: A new approach that converts a classification corpus into an sts one," p. 349–365, 2023.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.

[15] Y. Ma, Y. Wu, W. Su, Q. Ai, and Y. Liu, "Caseencoder: A knowledge-enhanced pre-trained model for legal case encoding," 2023.

[16] H. Li, Q. Ai, J. Chen, Q. Dong, Y. Wu, Y. Liu, C. Chen, and Q. Tian, "Sailer: Structure-aware pre-trained language model for legal case retrieval," 2023.

[17] Z. Zhou, J.-X. Shi, P.-X. Song, X.-W. Yang, Y.-X. Jin, L.-Z. Guo, and Y.-F. Li, "Lawgpt: A chinese legal knowledge-enhanced large language model," 2024.

[18] C. N. de Justiça, *Manual de Padronização de Ementas*, 2024. Acesso em: 25 maio 2025.

[19] A. Kwak, C. Morrison, D. Bambauer, and M. Surdeanu, "Classify first, and then extract: Prompt chaining technique for information extraction," pp. 303–317, 01 2024.

[20] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023.

[21] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. Le, and E. Chi, "Least-to-most prompting enables complex reasoning in large language models," 2023.

[22] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020.

[23] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," 2023.

[24] C. Singh, J. P. Inala, M. Galley, R. Caruana, and J. Gao, "Rethinking interpretability in the era of large language models," 2024.

[25] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," 2022.

[26] R. Egger and J. Yu, "A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts," *Frontiers in Sociology*, vol. Volume 7 - 2022, 2022.

[27] C. M. Pham, A. Hoyle, S. Sun, P. Resnik, and M. Iyyer, "Topicgpt: A prompt-based topic modeling framework," 2024.
[28] M. Wan, T. Safavi, S. K. Jauhar, Y. Kim, S. Counts, J. Neville, S. Suri, C. Shah, R. W. White, L. Yang, R. Andersen, G. Buscher, D. Joshi, and N. Rangan, "Tnt-llm: Text mining at scale with large language models," 2024.
[29] K. Didwania, D. Toshniwal, and A. Agarwal, "Unveiling themes in judicial proceedings: A cross-country study using topic modeling on legal documents from india and the uk," 2024.
[30] S. Aouichaty, Y. Maleh, M. T. Mohtadi, A. Hajami, and H. Allali, "Sustainable topic modeling for legal moroccan arabic language: A challenging study on bertopic technique," vol. 236, pp. 582–588, 2024. International Symposium on Green Technologies and Applications (ISGTA'2023).
[31] F. Souza, R. Nogueira, and R. Lotufo, "Bertimbau: Pretrained bert models for brazilian portuguese," (Berlin, Heidelberg), p. 403–417, Springer-Verlag, 2020.
[32] Serviço Federal de Processamento de Dados – SERPRO, "Serviço Federal de Processamento de Dados – SERPRO." https://www.serpro.gov.br/, 2025. Accessed: 2025-10-19.
[33] Mistral AI, "Apresentando o pixtral-12b — o novo modelo misto da mistral ai," 2024. Acesso em 24 mai. 2025.
[34] Conselho Nacional de Justiça, "Api pública datajud." https://datajud-wiki.cnj.jus.br/api-publica/, 2024. Acesso em: 9 ago. 2025.
[35] G. Bouma, "Normalized (pointwise) mutual information in collocation extraction," *Proceedings of the Biennial GSCL Conference 2009*, 01 2009.
[36] R. M. P. de Oliveira, "A perda de direitos gerados pelos contratos via mei, pejotização e a flexibilização da clt," 2023.
[37] Supremo Tribunal Federal, "Stf suspende processos em todo o país sobre licitude de contratos de prestação de serviços," 2024. Accessed: 2025-06-06.
[38] Academia Brasileira de Letras, "Abl na mídia – g1 – 'terrir', 'disania', 'pejotização'... 8 palavras podem entrar para o vocabulário oficial da língua portuguesa," July 2025. Acesso em: 20 jul. 2025.
[39] CGSIM, "Resolução cgsim nº 59, de 12 de agosto de 2020," 2020. Accessed: 2025-06-06.
[40] R. F. do Brasil, "Impedir/permitir participação do meu cpf no cadastro nacional de pessoas jurídicas (cnpj)," 2023. Accessed: 2025-06-06.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251574*

# Unconstitutionality Prevention in Bills Using Hybrid AI

Michele CORAZZA [a], Pier Francesco BRESCIANI [a], Generoso LONGO [a], Faria FEROOZ [a], Salvatore SAPIENZA [a] and Monica PALMIRANI [a,1]

[a] *Department of Legal Studies, CIRSFID - ALMA AI, University of Bologna*

ORCiD ID: Michele Corazza https://orcid.org/0000-0002-7288-6635, Pier Francesco Bresciani https://orcid.org/0000-0002-6889-1618, Generoso Longo https://orcid.org/0009-0003-2687-5884, Faria Ferooz https://orcid.org/0000-0002-8150-7852, Salvatore Sapienza https://orcid.org/0000-0002-5429-5217, Monica Palmirani https://orcid.org/0000-0002-8557-8084

**Abstract.** One of the most important legal assessments during legislative drafting proposals in Parliament is to check whether the bills have some provisions clashing with the Constitutional norms. Sometimes this situation is intentionally created by the proposer of a party to underline inconsistencies or incompleteness in the legal systems according to the reality of the society, but frequently, they are mistakes to avoid. This paper presents a method to calculate a multi-factor index of presumption of unconstitutionality of the provisions of a bill, to inform the proponent concerning a possible constitutional conflict. The paper uses a large dataset of bills, coming from the Chamber of Deputies of Italy, combined with Constitutional Court judgments and abstracts, EU and National legislation (9095 Italian laws, 3710 bills, 2979 judgments from the Constitutional Court). The methodology used is hybrid, considering different technological components, linked in a pipeline: i) embedding similarity; ii) normative references navigation; iii) semantic topic of the bill; iv) legislation cited in the judgments score based on the abrogated.

**Keywords.** Constitutional law, embedding, Akoma Ntoso, similarity, sentence transformer

## 1. Introduction and Research Questions

Constitutional review is a cornerstone activity of modern democratic governance aimed at ensuring that legislation aligns with the foundational principles enshrined in a nation's constitution. While *ex post* constitutionality review is usually demanded to specialised high-level courts, it is also carried out *ex ante* (i.e., before the bill is approved) by Parliaments on bills under discussion, to prevent the unconstitutionality of the forthcoming text. Such preliminary constitutionality review is traditionally carried out through legal scrutiny by parliamentary committees and designated departments of civil servants. However, the increasing complexity of legislative proposals and the evolving sociopolitical landscape suggest the possibility of integrating Artificial Intelligence (AI) systems in support of the parliamentary activities for this specific task. From a legal point of view,

[1] Corresponding Author: Monica Palmirani, monica.palmirani@unibo.it

this integration is not trivial. Principles of the rule of law and, in particular, constitutional law foundations [1] - e.g., separation of powers, the autonomy of the decision-maker [2], the law-making process - shall be conjugated with technical solutions that ensure accountability, fairness, and explainability of the outputs [3].

In this study, we investigate the feasibility of a novel automated assessment method of the constitutionality of bills while conjugating legal principles and the peculiarities of legal texts. We focus on the Italian legal system, a civil law country in which a Constitutional Court is tasked with an *ex post* assessment, and in particular on the activities of the Chamber of Deputies, one of the two branches of the Parliament, that has the obligation to check the possible unconstitutional provisions.

Given these premises, this study aims to investigate the following research questions:

- **RQ1**. How is it possible to represent the text of a bill with embedding while respecting its structure (paragraphs, subparagraphs, etc.) and navigating normative references point in time? This is a preliminary objective that will allow us to compare the bill with other material (e.g., decisions of the Constitutional Court).
- **RQ2**. How is it possible to fruitfully retrieve relevant decisions from the Italian Constitutional Court, given the text of a bill? This objective allows us to identify the decisions to compare with our bill.
- **RQ3**. How is it possible to compare a bill and the relevant decisions, and then to measure the risk of a bill being unconstitutional, given the bill and the correlated relevant decisions of the Constitutional Court?

To tackle these challenges, this study relies on a hybrid approach which combines an XML-structured representation of the legislative text in Akoma Ntoso and subsymbolic approaches leveraging on this structure [4]. Owing to the novelty and experimental nature of the approach, our results should be interpreted as preliminary, yet they are nonetheless encouraging. The paper is divided as follows: following this introduction, Section 2 introduces the background and the motivation underlying the study; Section 3 describes the datasets, whereas Section 4 details the methodology. Experimental settings are presented in Section 5, while their results are presented in Section 6 and discussed in Section 7. Some conclusive remarks are reported in Section 8.

## 2. Background and Motivation

In 2024, the 32% of the judgments of the Italian Constitutional Court were about the constitutional illegitimacy of law at the State and regional level (417 decisions, 132 of unconstitutionality) [2]. It is a problem of democracy, dialogue between institutions, and the division of powers because the Constitutional Court seems to invade the legislative space, especially in those domains with a lack of solid regulation (e.g., end of life). For this reason, the topic is urgent and very sensitive.

However, this task is not very well studied by the scientific research, that has concentrated mostly on prediction, factors classifications, summarization of the decisions.

Predicting courts decision is a classic task of the "AI & Law" domain [5,6]. Relevant studies in the field use NLP techniques in selected domain, and include notorious contributions in the European Court of Human Rights [7,8]. Among them, some studies focus on higher courts tasked with constitutionality assessment (e.g., "Supreme Courts",

[2] https://www.cortecostituzionale.it/annuario2024/pdf/Dati_quantitativi_e_di_analisi_2024.pdf Annual Report 2024

"High Court", and so forth). This has been the case for studies focusing on the United States [9], Brasil [10], Germany [11], Singapore [12], and many other countries.

Surveys show a consistent pattern of duality: on the one hand, certain studies focus on case law retrieval from higher courts and have the goal of retrieving pertinent to a set of facts (in most cases) or a given piece of legislation [13]; on the other hand, some studies focus on the constitutionality assessment as a classification task in which the study's goal is to predict the outcome of the Court's decision given a set of facts or prior cases [14]. This can happen with or without the case retrieval. Large Language Models (LLM) are showing advances in both fields [15,16,17], despite the chance of hallucinations[18] and technical constraints (e.g., context length).

Our study differs from the existing literature on three grounds:

- it focuses on the Italian legislative system, which has not been subject to particular attention in the domain of unconstitutionality evaluation. This might be due to several reasons, including data retrieval and language complexity;
- it processes bills rather than in-force legislation or sets of facts. While training datasets of the existing literature use applicable legislation, we directly measure the risk of unconstitutionality over bills. The difference lies on the fact that in-force laws are already interpreted by lower courts and constitutional courts have prior interpretations at their disposal. From a legal standpoint, assessing the constitutionality of bills is harder when compared to applicable legislation;
- it uses an XML-enriched version of the legal texts to allow the point-in-time navigation of legal references, rather than plain text. This contributes to increase the proximity between the reasoning performed by the automated system and the one performed by the Constitutional Court.

## 3. Datasets

In our experiments, we used three distinct datasets from different legal resources: i) Italian legislation; ii) bills; iii) Constitutional Court decisions. Crucially, all these dataset are provided in Akoma Ntoso XML or they were converted by us for our experiments. The three datasets that we used are:

- **Italian legislation in Normattiva**: a dataset of Italian legislative documents extracted from the official website Normattiva containing Italian laws. It is composed of all documents from 2000 to 2025, 9095 in total, including all consolidated versions. Furthermore, we extracted information about the abrogated articles of these documents, which are used in our user interface to show judgments that include normative references to abrogated norms. This parameter is fundamental for informing legal experts regarding the up-to-date status of the legislative information.
- **Chamber of Deputies Bill (PDL)**: a dataset of 3710 legislative bills from the XVIII and XIX (March 2018 to May 2024) legislatures. They were downloaded from the official website of the Italian Chamber of Deputies and converted to Akoma Ntoso XML using an ad-hoc procedure[3]. In the paper now we mention this dataset simply as CD Italian bills.
- **Italian Constitutional Court Decisions**: our starting point is a dataset containing all the judgments from the Italian Constitutional Court converted to Akoma Ntoso. We first extracted the more recent ones (from 2001 to 2025). Then, the metadata

[3] https://gitlab.com/CIRSFID/html2aknPDL

| Result (Italian) | English translation | Annotation |
|---|---|---|
| ill. cost. conseg. ex art. 27 legge n. 87/1953 | consequential uncostitutionality ex Art. 27 Law No. 87/1953 | negative |
| ill. cost. parziale conseg. ex art. 27 legge n. 87/1953 | consequential partial uncostitutionality ex Art. 27 Law No. 87/1953 | negative |
| illegittimità costituzionale | unconstitutionality | negative |
| illegittimità costituzionale parziale | partial unconstituitonality | negative |
| manifesta infondatezza | manifest unfoundedness | positive |
| non fondatezza | unfoundedness | positive |

**Table 1.** The relevant results of the judgments and their annotation

associated with each document was used to extract judgments that contain at least one verdict of unconstitutionality, obtaining 2979 judgments. The full list of verdicts, as they are shown in the metadata, can be found in Table 2.

In addition to selecting the verdicts, it was also crucial to select the verdict results and their relative annotation as either positive (non-unconstitutionality was found) or negative (verdict of unconstitutionality). These result types, as shown in the metadata, are found in Table 4.

## 4. Methodology

Our approach is to build a meaningful index of unconstitutionality for CD Italian bills according to legal principles and considering the similarity of the bill with the relevant decisions of the Constitutional Court that have already decided unconstitutionality. While an alternative approach could have lead to find similarities between the current bill and legislation already scrutinised by the Court, this would have lead to the inclusion of legislation already assessed and, in turn, some degree of bias towards the prediction. Therefore, we have considered three parameters: the similar domain, the similar legal concepts, and the similar rationale in the legal reasoning.

In order to produce a meaningful measure of the risk of unconstitutionality, the first step is to match bills with the relevant judgments from the Constitutional Court. One might be tempted to use a lexical comparison based on TF-IDF or similar techniques. These approaches, however, rely on the assumption that the lexical overlap between the two documents being compared is significant. In the case of the comparison between bills and constitutional court judgments, these are very different types of documents, which have a different structure an lexicon. For this reason, our approach is based on a sentence transformer [19], a type of model which allows the creation of embeddings for sequences of text, which can then be used to find semantically similar texts. This is typically achieved by using a contrastive learning loss, which forces the vectors for similar sentences to have a high cosine similarity.

While sentence embeddings can be used to compare the semantics of relatively short portions of text, the ability of any model to deal with longer documents is not guaranteed. In fact, even LLMs struggle with documents, recent results show that the effective length of most documents is lower than the maximum number of tokens by at least an order of magnitude [20] and they are effectively able to deal with less than 10k tokens. To produce a meaningful semantic similarity metric between Italian bills and Italian Constitutional Court judgments, we considered the following portions of these documents:

- **CD Italian bills**: by leveraging the Akoma Ntoso format we are able to access the

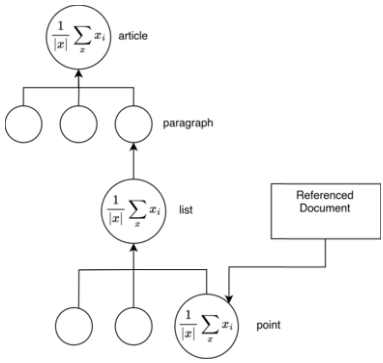



**Figure 1.** A graphical visualization of a possible application of the tree aggregation strategy. In this case, an article contains three paragraphs, and the last one contains a list of three points, the last point contains a normative reference.

structure of Italian bills and we operate at the granularity of a single article.

- **Chamber of deputies commission**: we use an unsupervised approach to associate each bill with the relevant Italian Chamber of Deputies commission. This is fundamental in order to detect the legal topics of the bill from the legislative competencies of the committee.
- **Constitutional Court decisions**: we select the paragraphs that are associated with the motivation of the judgment, using the corresponding tag of AKN `<motivation>`.
- **Constitutional Court abstracts**: these are abstracts (massime) which are redacted directly from the Constitutional Court. There can be more than one associated with each decision, as each one may contain more than one interpretation if the constitutionality assessment spans across different provisions.
- **Constitutional Court *ratio decidendi***: a series of terms are usually associated by the judges to the decision, remarking the main legal reasons which are associated with each judgment directly from the court.

The discussion of how these different textual sources were used in our approach can be found in the next sections.

### *4.1. Tree embeddings*

Both bills and decisions from the Constitutional Courts, however, have a hierarchical structure (bill has articles, commas, paragraphs, etc. and decisions introduction, background, motivation, decision) and they can contain legal citations. For this reason, we adopt a tree aggregation strategy for our embeddings. For the sake of clarity, we describe this approach as it is applied to articles of CD Italian bills. For the paragraphs of the motivation of the Italian Constitutional Court decisions, we use the same procedure, but we treat the references to other judgments in a slightly altered manner.

Given a sentence transformer model, instead of applying it directly to the text of the entire article, which might be too long for the maximum number of tokens for sentence transformer models, we traverse its Akoma Ntoso XML tree, until we find a leaf node. These leaf nodes were defined manually and they are based on a list of XML elements that are marked as inline in the Akoma Ntoso standard (references, dates, modifications,

etc). Our algorithm proceeds from the article to all its children nodes that do not belong to this list. Then, we need to consider normative references. These references can be punctual (reference to a single article, point, section, etc) or they can be generic references (to an entire document). Formally, given an article $a$ which contains children and references, and a sententce transformer model $M$, we use the following recursive function:

$$v(a) = \frac{1}{2+|c(a)|}\left(M(t(a)) + \sum_{i=1}^{|c(a)|} v(c_i(a)) + \frac{1}{|r(a)|}\sum_{h=1}^{|r(a)|} R(r_h(a))\right) \tag{1}$$

Where:

- $t(a)$ is the text contained in the article which is not a part of any of its children
- $c_i(a)$ is the i-th non inline child of the article
- $c(a)$ is the set of all the non inline children of the article $a$
- $r_h(a)$ is the h-th normative references which is directly contained in $t(a)$
- $r(a)$ is the set of all normative references which are directly contained in $t(a)$

In order to represent normative references, we define the function R as follows:

$$R(i) = \begin{cases} v(i) & \text{if } i \text{ is a punctual reference} \\ \frac{1}{2}M(T(i)) + v(A_1(i)) & \text{if } i \text{ is a generic normative reference} \\ \frac{1}{2}M(I(i)) + v(P(i)) & \text{otherwise} \end{cases} \tag{2}$$

Where:

- $T(i)$ represents the title of the referenced document
- $A_1(i)$ represent the first article of the referenced document
- $T(i)$ represents the introduction of the referenced judgment
- $P(i)$ represent the motivation of the referenced document

By adopting these recursive functions, we obtain a procedure which creates a vector representation for each article by leveraging its structure, and which also considers the normative references contained in the documents.

### 4.2. Bill-Decisions Similarities

In order to produce a measure of similarity between a bill and the existing decisions from the Constitutional Court, we use three different similarity measures.

#### 4.2.1. Text Similarity

The first measure is the average similarity between each article of a bill and each paragraph from the `<motivation>` tag of the decision. Formally, given a bill $b$ and a decision $j$, the similarity is measured as:

$$s_{ap}(b,j) = \frac{1}{|A(b)||P(j)|}\sum_{i=1}^{|A(b)|}\sum_{h=1}^{|P(j)|} sim(v(A_i(b)), v(P_h(J))) \tag{3}$$

Where:

- *sim* represents cosine similarity;

- $A(b)$ represents all the articles of bill $b$, $A_i(b)$ is the i-th article of bill $b$
- $P(j)$ represents all the paragraphs of the motivation of judgment $j$, $P_h(j)$ is the h-th paragraph of the motivation of judgment $j$

#### 4.2.2. Legal Concepts Similarity

Another criterion for the comparison of the judgment and the bill is based on the *ratio decidendi* of the judgment (which qualifies the final result of the decision) and the Italian Chamber Commission assigned to the bill (that qualifies the topics and competencies of the bill). Since most of the bills are never assigned a Commission, we first use an unsupervised approach based on similarity to assign the Commission to each bill $b$ as follows:

$$Z(b) = argmax_h \left( \frac{1}{|A(b)|} \sum_{i=1}^{|A(b)|} sim(v(A_i(b)), M(k_h)) \right) \tag{4}$$

Where:

- $k$ represents all the Italian Chamber commissions;
- $k_h$ is the textual description of the h-th commission, as found in the Chamber regulation as provided by its implementation in accordance to art. 22, paragraph 1-bis Lettera circolare sugli ambiti di competenza delle Commissioni permanenti, 16 October 1996;

Once we obtained the Commission assigned to each bill, the similarity is calculated as follows:

$$s_{cz}(b, j) = sim(M(k_{Z(b)}), M(d(j))) \tag{5}$$

Where $d(j)$ is the ratio decidendi associated with the current judgment $j$.

#### 4.2.3. Reasoning Similarity

Finally, another similarity criterion is the one involving Judgment abstracts and Articles of the bill. Formally, given a bill $b$ and a judgment $j$, this similarity can be expressed as:

$$s_{am}(b, j) = max_h \left( \frac{1}{|A|} \sum_{i=1}^{|A(b)|} sim(v(A_i(b)), M(m_h(j))) \right) \tag{6}$$

Where $m_h(j)$ represents the h-th abstract associated with judgment $j$. In this case, since one judgment can be associated with multiple abstracts, we opted to use the maximum similarity between the articles of the bill and each of the abstracts. This is motivated by the possibility of one of the verdicts being more relevant for the bill than the others.

### 4.3. Unconstitutionality prevention index

In order to show the user a measure of risk for unconstitutionality for a given bill, we first select the 10 most similar judgments from the court using one of the similarity metrics from the previous section. Then, the judgments that contain at least one verdict of unconstitutionality are considered, and the similarity value associated with them is

used to scale their relative importance in the risk formula. Formally:

$$R(b) = \frac{\sum_{h=1}^{10} s_h(b) I_h(b)}{\sum_{i=1}^{10} s_i(b)} \tag{7}$$

Where $s_i$ is the i-th similarity value in the 10 most similar judgment to bill $b$, while $I_h(b)$ is defined as follows:

$$I_h(b) = \begin{cases} 1 & \text{if the h-th judgment contains at least an unconstitutionality verdict} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

By adopting this method, we produce a risk value which is influenced by the relative similarity of the judgments to the bill and which is able to provide the stakeholders with a meaningful metric for the risk of unconstitutionality associated with each bill.

| Italian | English translation |
|---|---|
| Giudizio di legittimità costituzionale in via incidentale | Incidental controversies on the constitutional legitimacy of laws |
| Giudizio di legittimità costituzionale in via principale | Direct controversies on the constitutional legitimacy of laws |
| Questione incidentale di legittimità costituzionale | Incidental question of constitutional legitimacy |

**Table 2.** The types of judgments which are relevant for our analysis

## 5. Experimental Settings

From the technical point of view, our goal is to compare the three similarity metrics discussed in Section 4 and to assess the overall feasibility of this approach. We opted to use one of the multilingual models from SentenceBERT, namely "paraphrase-multilingual-mpnet-base-v2" [21]. This model is trained from a monolingual Sentence Transformer model, namely "paraphrase-mpnet-base-v2" which is based on MPNet [22]. The adopted approach is based on the teacher-student paradigm, where the goal is to produce a model where the original and translated versions of the same sentence have a high cosine similarity.

In order to assess the performance of our approach, we tasked an expert in constitutional law to produce a list of 10 bills that involve different topics. Then, the expert performed a preliminary analysis by annotating relevant decisions that could be used to assess the risk of unconstitutionality for each bill. In particular, the focus was the creation of a list of subjects that potentially involve matters of constitutionality, and a list of judgments that are the most important precedents for each subject. Then, the expert manually checked the 10 most similar judgments to each selected bill, as produced by the three similarity metrics. Finally, the 10 most similar judgments were annotated as either "relevant" or "not relevant".

The selected bills (Table 3) were identified by our Constitutional Law expert on the basis of recent Constitutional Court case-law on the topic. Furthermore, to provide a representative sample of different types of bills, both bills that only modify existing laws and new legislation (which might also involve modifications of existing laws) were considered.

| Topic | Bill | Type of legislative intervention |
|---|---|---|
| End of life | Bill No. 313/2022 | New Act |
| Employment rights on termination | Bill No. 277/2018 | Amending Act |
| Term limits for elected offices | Bill No. 1644/2024 | Amending Act |
| Regional autonomy | Bill No. 1665/2024 | New Act |
| Protection of animals | Bill No. 1109/2023 | New Act |
| Protection of cultural heritage | Bill No. 1297/2023 | New Act |
| Life without parole | Bill No. 1951/2019 | Amending Act |
| Children's surname | Bill No. 256/2022 | Amending Act |
| Adoption by single persons | Bill No. 467/2018 | Amending Act |
| Criminal penalties for drug offences | Bill No. 71/2022 | Amending Act |

**Table 3.** Type of topics chosen for the experiment

| Similarity Criterion | Top 10 Recall | Top 5 Recall |
|---|---|---|
| Articles-Paragraphs | 0 | 0 |
| Commission-Ratio decidendi | 0 | 0 |
| Articles-Abstract | 0.51 | 0.31 |

**Table 4.** Results of the experiments with 10 bills

## 6. Results Evaluation

To measure the performance of the various similarity criteria, we used recall as a metric, since we have access to the ground truth judgments provided by our expert, as well as those that are within the top 10 and top 5most similar judgments to a given bill. This choice of 10 and 5 judgments was motivated by a mediation between providing users with as many relevant results as possible, while not overwhelming them with too many, which would make the method less useful. The results of our evaluation are shown in Table 4 and they show the recall of the system when using the top 10 and top 5 most similar judgments, respectively. Since we do not have access of a complete list of true positives, as this would require an extensive and time-consuming exploration of all the constitutional court judgments, we are unable to calculate the precision of the system. In our experiments, the comparison between bill articles and the paragraphs of the motivation proved unsuccessful. This can perhaps be explained by the fact that this comparison involves many different portions of text, which makes it less reliable. Furthermore, the differences in language between a judgment and a bill might also have impacted the performance. The comparison between Chamber of Deputies Commissions and *ratio decidendi* from the Court also proved unsuccessful. This result might be due to the fact that the scope of each Commission is very broad, meaning that some details about the specific bills cannot be represented correctly. Finally, our best results come from the comparison between bill articles and abstracts from the Constitutional Court, which is probably due to the more concise and simpler language used in the abstract when they are compared with the overall judgments. Furthermore, since each verdict is associated with its own abstract, this further enhances the granularity of the approach, allowing the embeddings to be more accurate.

## 7. Discussion

In order to deepen our understanding of the shortcomings of our approach, we conducted a post hoc legal analysis of the results. Firstly, we observed that in two cases (Term Limits for Elected Offices and Criminal Penalties for Drug Offences), our system was unable to retrieve any relevant judgments. From a legal perspective, both bills made only very

specific modifications to highly technical existing legislation. Specifically, the bill on term limits for elected offices consisted of a single substantive article and contained only one explicit reference to the area of legislation affected by the modification. This suggests that our methodology performs better when the theme of the legislation is repeatedly expressed within the legislative texts, whereas it tends to struggle when the context must be inferred primarily through normative references. Secondly, in the case of the bill on children's surnames, the system was able to retrieve all the relevant case-law annotated by our expert. This bill, while modificative in nature, amends the Civil Code, which is a particularly structured type of legislation, with a set of topics that are generally mutually related and with semantically informative article titles. Thirdly, when comparing true positives and false negatives, it can be observed that the approach struggles to identify the most recent relevant judgments. In many cases (e.g., end-of-life or employment rights on termination), the system identified some relevant case-law but missed the most recent judgments on the matter. This limit is legally significant because of the evolving nature of constitutional case-law, which requires constitutionality analyses to begin with the judgments that best reflect the approach of the sitting court. Finally, the system suggested judgments that were not annotated ex ante by our expert but were considered relevant in the post hoc analysis. This highlights that very complex, time-consuming, and open-ended legal tasks, such as researching the relevant case-law to assess the constitutionality of bills, can benefit from assistive AI-based applications, which are capable of processing large volumes of information in a short amount of time.

## 8. Conclusive Remarks

This paper demonstrates an innovative hybrid pipeline to calculate an index of unconstitutionality of the bills of the Chamber of Deputies of Italy. We have fostered the AKN-XML annotation of the heterogeneous legal sources repository, using embeddings to prepare vectors for similarity comparison. We have also combined different levels of similarity to grasp different aspects of the legal principles in the Constitutional Law to support the unconstitutionality analysis. The results are promising, but we intend to add more levels of legal analysis in the future. We aim to improve our work, including Question/Answering knowledge graphs coming from the legal knowledge of domain. Secondly, we intend to annotate some *ratio decidendi* with an ontology in order to improve the embeddings with RAG based on supervised legal knowledge. Thirdly, we will start an in-place evaluation made by legal experts of the Chamber of Deputies to capture more reasoning models to inject in our system. Another improvement could be a time-sensitive modeling of the unconstitutionality index, which would consider crucial temporal aspects such as the weight of abrogated norms. Finally, we intend to further improve and validate our results by including articles of the Constitution in the analysis.

## Acknowledgments

This project is conducted with the support of the European Commission funds within ERC HyperModeLex. Grant agreement ID: 101055185. This project is co-funded by the European Union - NextGenerationEU under the National Recovery and Resilience Plan (PNRR) - Mission 4 Education and research - Component 2 From research to business - Investment 1.1 Notice PRIN 2022 (DD N. 104 del 02/02/2022), title Smart Legal Order in DigiTal Society (SLOTS), proposal code 2022LRL2C2 - CUP J53D23005610006.

## References


[1] Pier Francesco Bresciani MP. Constitutional Opportunities and Risks of AI in the law-making process. Federalismie. 2025;2/2024.

[2] Pier Francesco Bresciani MP. AI Act and National Parliaments: How the EU is Regulating Democracy in the Digital Era. Ragion pratica, Rivista semestrale. 2025;1/2025:249-68.

[3] Palmirani M. A smart legal order for the digital era: a hybrid AI and dialogic model. Ragion pratica. 2022;(2):633-55.

[4] Palmirani M, Sapienza S, Ashley K, et al. A hybrid artificial intelligence methodology for legal analysis. Biolaw Journal. 2024;3:389-409.

[5] Ashley KD. Artificial intelligence and legal analytics: new tools for law practice in the digital age. Cambridge University Press; 2017.

[6] Ashley KD. A brief history of the changing roles of case prediction in AI and law. Law Context: A Socio-Legal J. 2019;36:93.

[7] Tsarapatsanis D, Aletras N, Preoţiuc-Pietro D, Lampos V. Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. PeerJ Computer Science. 2016 Oct.

[8] Medvedeva M, Vols M, Wieling M. Using machine learning to predict decisions of the European Court of Human Rights. Artificial Intelligence and Law. 2020;28(2):237-66.

[9] Katz DM, Bommarito II MJ, Blackman J. Predicting the behavior of the supreme court of the united states: A general approach. arXiv preprint arXiv:14076333. 2014.

[10] Bertalan VGF, Ruiz EES. Predicting Judicial Outcomes in the Brazilian Legal System Using Textual Features. In: DHandNLP@ PROPOR; 2020. p. 22-32.

[11] Waltl B, Bonczek G, Scepankova E, Landthaler J, Matthes F. Predicting the outcome of appeal decisions in germany's tax law. In: International conference on electronic participation. Springer; 2017. p. 89-99.

[12] SOH JTH, CHAI IE. Legal area classification: A comparative study of text classifiers on Singapore Supreme Court judgments.(2019). In: Proceedings of the Natural Legal Language Processing Workshop 2019, Minneapolis, Minesotta, June. vol. 7;. p. 67-77.

[13] Feng Y, Li C, Ng V. Legal case retrieval: A survey of the state of the art. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2024. p. 6472-85.

[14] Alcántara Francia OA, Nunez-del Prado M, Alatrista-Salas H. Survey of text mining techniques applied to judicial decisions prediction. Applied Sciences. 2022;12(20):10200.

[15] Kim C, Lee J, Hwang W. LegalSearchLM: Rethinking Legal Case Retrieval as Legal Elements Generation. arXiv preprint arXiv:250523832. 2025.

[16] Deng C, Mao K, Dou Z. Learning interpretable legal case retrieval via knowledge-guided case reformulation. arXiv preprint arXiv:240619760. 2024.

[17] Bertalan VGF, Ruiz EES. Using attention methods to predict judicial outcomes. Artificial Intelligence and Law. 2024;32(1):87-115.

[18] Magesh V, Surani F, Dahl M, Suzgun M, Manning CD, Ho DE. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. Journal of Empirical Legal Studies. 2025;22(2):216-42.

[19] Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2019. Available from: https://arxiv.org/abs/1908.10084.

[20] Modarressi A, Deilamsalehy H, Dernoncourt F, Bui T, Rossi RA, Yoon S, et al. NoLiMa: Long-Context Evaluation Beyond Literal Matching. In: Forty-second International Conference on Machine Learning; 2025. Available from: https://openreview.net/forum?id=0OshX1hiSa.

[21] Reimers N, Gurevych I. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2020. Available from: https://arxiv.org/abs/2004.09813.

[22] Song K, Tan X, Qin T, Lu J, Liu TY. Mpnet: Masked and permuted pre-training for language understanding. Advances in neural information processing systems. 2020;33:16857-67.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*

*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251575*

# Automated Creation of a Legal Knowledge Graph Addressing Cases of Violence Against Women

*Resource, Methodology and Lessons Learned*

Claudia D'AMATO [a] and Giuseppe RUBINI [a] and Francesco DIDIO [a] and Donato FRANCIOSO [a] and Fatima Zahra AMARA [a] , and Nicola FANIZZI [a]
[a] *Computer Science Department, University of Bari Aldo Moro, Italy*
ORCiD ID: Claudia d'Amato https://orcid.org/0000-0002-3385-987X, Fatima Zahra Amara https://orcid.org/0000-0001-8463-0330, Nicola Fanizzi https://orcid.org/0000-0001-5319-7933

**Abstract.** Legal decision-making requires comprehensive legislative knowledge and up-to-date case information. Legal Knowledge Graphs can enhance accessibility of such information, support semantic querying, and serve as knowledge-intensive components for predictive machine learning applications. To address the limited availability of legal KGs, this work develops a KG focused on cases of violence against women. Two complementary construction approaches are presented: a bottom-up, domain-customized methodology and a novel Large Language Model-based solution. Both integrate data extraction, ontology development, and semantic enrichment using sentences from the European Court of Justice as sources. The resulting KGs are validated through competency questions and show potential to improve access to legal information, enable complex queries, and provide a valuable knowledge backbone for supporting predictive tasks such as case outcome analysis.

**Keywords.** Knowledge Graph Generation, Large Language Models, Linked Data, Semantic Web, Violence Against Women, Prediction.

## 1. Introduction

In recent years, the legal domain has witnessed an increasing usage of solutions based on Artificial Intelligence (AI) to effectively manage the legal consultation and decision process [1,2]. Despite the advancement of legal technologies [3], there is a significant gap in the availability of structured and easily queryable resources tailored to the legal domain, such as Knowledge Graphs (KGs) [4]. Current KGs in the legal domain often do not address specific subtle challenges, particularly with regard to interoperability, semantic richness, and adaptability to predictive tasks [5,6].

A legal KG must be able to cope with the specific and intricate writing style of laws and sentences and keep track of their evolution, including exceptions and applicability.

We address the urgent global phenomenon of violence against women[2], a violation of human rights with serious physical, mental, economic and social consequences [7]. Effectively countering this problem requires coordinated actions from governments, civil society, and individuals. Our work contributes by developing advanced tools and knowledge resources, particularly knowledge graphs, to support the analysis of judicial reasoning, facilitate policy evaluation, and improve access to structured case-law data, on gender-based violence. To this end, we introduce a novel Legal KG, specifically designed to address judicial cases related to violence against women, along with the methodologies adopted for its development.

The Legal KG is automatically derived from the jurisprudence of sentences of the *European Court of Human Rights*[3] (ECHR), targeting cases of gender-based violence. It adheres to the FAIR (*Findable*, *Accessible*, *Interoperable*, *Reusable*) data principles. The resource is built upon two methodologies: a customization of the general KG development process to the legal domain [8] (with a bottom-up solution combining data extraction, ontology development, and semantic enrichment through alignment to domain-specific ontologies), and a new approach to automated KG generation, based on the exploitation of *Large Language Models* (LLMs), as recent advancements have opened up new possibilities for automating and improving various phases of ontology development[9], that could be customized to the legal domain. The two methodologies produced complementary results: the former provided more precise outcomes, but is more computationally expensive while the latter is more scalable, but ensured less accuracy. The obtained resource has also been validated via suitable CQs (*Competency Questions*) [10].

Besides being the first result addressing the ECHR case law to cases of violence against women, the Legal KG may serve as a proxy for broader applications. Judicial pronouncements prediction can be thus viewed a link prediction problem within the Legal KG. This approach can be generalized to other domains, provided that legal KGs are generated through the proposed methodologies.

The remainder of this paper is structured as follows: Sect. 2 reviews the state of the art on legal informatics, KGs and ontologies in law. Sect. 3 presents a customization to the legal domain of the bottom-up KG development process and the resulting Legal KG. Sect. 4 presents a new approach for automated KG generation, grounded on the exploitation of LLMs and customized to the legal domain, and the obtained Legal KG. Sect. 5 discusses the results and the effectiveness of the adopted methodologies. Sect. 6 concludes the paper by outlining future research directions.

## 2. Related Work

This section reviews state-of-the-art work in legal informatics with a focus on KGs, ontologies, and thesaurus initiatives, covering methods for structuring legal data, automated reasoning, and extracting insights from texts.

The *Joint Knowledge Enhancement Model* refines LLMs for improved extraction, applied in building the *Chinese Legal KG* with 3480 triples but no domain ontology reuse [2]. Similarly, a judicial case KG leveraging joint event extraction and Dgraph en-

[2] https://www.who.int/news-room/fact-sheets/detail/violence-against-women
[3] https://www.echr.coe.int/

ables retrieval and case analysis but lacks interlinking [11]. The ManyLaws platform provides advanced legal data access in the EU through analytics and visualization, though without semantic querying [12]. KG-based approaches for automatic document generation improve efficiency and quality but do not enable semantic queries [13].

At the operational level, ontologies support compliance checking by modeling deontic norms and enabling reusable reasoning patterns [14]. An RDF-based graph for fine-grained document retrieval leverages ontologies but requires heavy manual construction [15].

Prominent KG efforts include the Lynx project on GDPR and contract compliance, structuring legal documents and linguistic data [16]. Inspired works process institutional publications for queryability with limited ontology reuse [17], while Vietnamese and other KGs [18,19] similarly show ontology underuse, though some incorporate ontology design patterns.

Finally, ontologies and thesauri have been developed to standardize legal terminology and foster interoperability across jurisdictions [20]:

**EuroVoc**[4] is a multidomain, multilingual thesaurus provided by the *Publications Office* of the European Union used to classify EU documents into categories to facilitate information searching. It is based on the *Simple Knowledge Organization System*[5] (SKOS) standard that is used to represent and organize concepts and vocabularies.

**European Law Identifier**[6] is a standard created to identify legislative documents from European states by providing an ontology and metadata. It provides simplified access, exchange and reuse of legislation for legal professionals or citizens, and forms the basis for a representation of the Official Journals of the member states in the Semantic Web[7].

**European Case Law Identifier**[8] **(ECLI)** defines a standard identifier for European jurisprudence, together with a minimal set of metadata. The identifier provided by ECLI is intended to have an identification code whose structure is common among all member states. This is divided into five parts separated by the colon as in the following example: `ECLI:CE:ECHR:2022:0210JUD007397516`.
ECLI identifies case-law and ELI legislative texts, serving complementary roles. To our knowledge, this is the first KG resource dedicated to legal cases of violence against women, providing SPARQL queryability, FAIR compliance, and integration with existing ontologies and legal thesauri, thereby addressing European legislative specifics and supporting broader legal reasoning and decision-making.

## 3. Legal Knowledge Graph Construction: Bottom-Up Approach

Advancements in KG engineering have significantly impacted the legal domain by supporting the organization of legislative texts, case law, and regulations, thereby improving legal reasoning, prediction, data integration, and complex querying. Following the

[5] https://www.w3.org/TR/skos-reference/

[7] Major progress towards transparency: a new European legislation identifier: https://ec.europa.eu/commission/presscorner/detail/en/IP_12_1040.

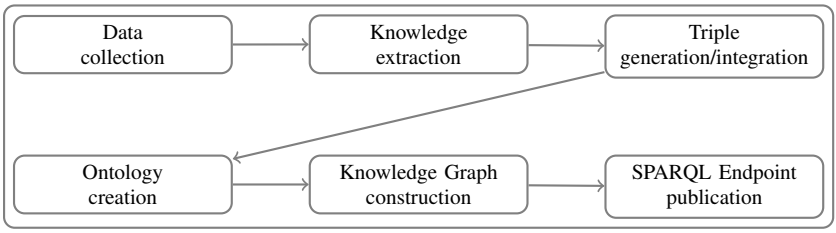



**Figure 1.** The adopted pipeline.

general methodology proposed by Tamašauskaitė and Groth [8], which distinguishes between top-down and bottom-up approaches, we adapt the bottom-up strategy to cases of violence against women. The process consists of six steps: data collection, knowledge extraction, processing and triple generation, ontology creation, KG construction, and maintenance, with the resulting KG accessible via a SPARQL endpoint. The full pipeline is openly available as a Python project with documentation on GitHub[9].

Regarding **Data collection**, ECHR judgments and decisions, accessible through the official website[10] have been collected and referenced in a compiled document[11]. Specifically, 73 judgments and decisions (in English[12]) have been selected by experts in international law. For each judgement/decision, we specifically accessed tabs `View` and `Case Details` where `View` tab provides the full pdf document, while `Case Details` tab contains additional semi-structured information (in HTML format), such as publication details, importance level, and ECLI identifiers, in agreement with the specific format: `ECLI:EC:ECHR:year`. For example, `ECLI:EC:ECHR:2022:0210JUD007397516` denotes a judgment from February 10, 2022, with appeal number `73975/16`.

Using Selenium 4 Library[13], we set up a functionality[14] for automatic download PDF and HTML files for each url sentence, and have them ready for the successive knowledge extraction step. Indeed, Selenium allows to control a web browser automatically and simulate human actions, such as clicking with the mouse on a specific item.

The **Knowledge Extraction** and **Triple Generation** phases process sentence files to extract triples, using HTML files as the main source. Each sentence instantiates an `ECHRDocument` class. The Beautiful Soup library[15] was employed to extract information, including converting defendant state names to Wikidata URIs and standardizing Importance Level values. From 73 selected judgments (65) and decisions (8), a total of 10,325 triples were extracted, involving 22 predicates and 5,185 entities. These triples were serialized into JSON or processed with RDFLib[16] to produce RDF/Turtle files. Redundancy checks confirmed no duplicates within sentences, while RDFLib automatically removed cross-file duplicates.

[9] https://github.com/PeppeRubini/EVA-KG
[10] https://hudoc.echr.coe.int/
[11] https://github.com/PeppeRubini/EVA-KG/blob/main/data/mapping_doc_link.xlsx
[12] https://hudoc.echr.coe.int/eng
[13] https://selenium.dev/
[14] https://github.com/PeppeRubini/EVA-KG/blob/main/src/echr_scraper.py
[15] https://beautiful-soup-4.readthedocs.io/en/latest/
[16] https://rdflib.readthedocs.io/en/stable/gettingstarted.html

As for the **Ontology Creation** and **KG construction**, in agreement with the ontology engineering best practices, we started with the formalization of CQs[17] that the ontology-based KG should address. The CQs adopted for this purpose are summarized in Tab. 1 (first column). For this phase, we also considered existing ontologies and vocabulary, such as ECLI, for possible reuse for representing ECHR pronouncements. However, not all CQs could have been answered using the existing resources (Tab. 1, second column). New concepts such as `DomesticLaw`, `InternationalLaw`, and `StrasbourgCaseLaw` were introduced, along with properties including `applicationNumbers`, `importanceLevel`, `respondentStates`, `involvedArticles`, and `unanimousDecisionIndicators`. Individuals were created for each ruling under `StrasbourgCaseLaw`, using URI sentences. Overall, metadata from Case Details enriched the ECLI data and addressed most CQs.

The final KG has been constructed by combining the triples from various judgments and linked to external resources including Wikidata for geographic information representation. We also developed three independent visualization solutions by adopting: PyVis Library[18], RDF Grapher[19], and Neo4j[20], respectively.

In order to query the obtained KG, we proceeded with the **SPARQL Endpoint publication**. Specifically, the SPARQL endpoint has been implemented as a Flask[21] web application, leveraging the flexibility and simplicity of the Flask framework for web services. Flask, serves as the web framework for handling HTTP requests, routing, and rendering responses. SPARQL 1.1 engine has been used for executing the SPARQL queries. RDF data is loaded into the RDFLib graph at application startup. The graph supports multiple serialization formats, such as Turtle, RDF/XML, and N-Triples. The application includes two main routes: the home page (`/`) with a form for submitting queries and a query endpoint (`/query`) that processes and displays the results. The interface uses HTML templates: `INDEX.html` for query submission and `RESULT.html` for displaying results. This setup provides a user-friendly way to interact with and query RDF data. The SPARQL endpoint, along with the full implementation of the web application, is available on the GitHub repository[22].

A KG on violence against women was built from 73 European Court of Human Rights judgments, containing 5,185 triples with 22 nodes, 1,747 properties, 5,108 subjects, and 10,325 objects this describes the cleaned KG supporting CQs on document details (type, date, parties, laws) and links to Wikidata for geographic information. The resulting KG, publicly available[23], provides a semantically rich interconnected structure that enables advanced querying and analysis of ECHR decisions pertaining to gender-based violence. It leverages established semantic web standards and vocabularies, enabling interoperability with external resources and datasets. The dataset has been published as part of the *Linked Data* cloud[24], adhering to the FAIR principles (Sect. 5).

[17] https://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html

[18] https://pypi.org/project/pyvis/

[19] https://www.ldf.fi/service/rdf-grapher

[20] https://neo4j.com/

[21] https://flask.palletsprojects.com/en/3.0.x/

[22] https://github.com/khaoulafatima/PJ4W

[23] https://github.com/PeppeRubini/EVA-KG/blob/main/KG.ttl

[24] The dataset can be found at: https://lod-cloud.net/dataset/PREJUST4WOMAN_PROJECT

**Table 1.** Competency Questions Adopted for the Bottom-Up Approach.

| Question | Corresponding Term |
|---|---|
| **What type of document are we dealing with?** | `dcterms:type` |
| **When is the document dated?** | `dcterms:date` |
| **Can I retrieve the information of a certain case given its identifier?** | `dcterms:isVersionOf` |
| **Who represents the applicant?** | `dcterms:contributor` |
| **To which European state does the applicant belong?** | `respondentState` |
| **What was the ruling?** | `dcterms:abstract` |
| **Was the ruling unanimous?** | `unanimousDecision` |
| **Which articles of the Convention were considered?** | `involveConventionArticle` |
| **Which laws were considered to reach this conclusion?** | `dcterms:references` |
| **What is the importance of the ruling concerning future cases?** | `importanceLevel` |
| **Is the document publicly accessible?** | `dcterms:accessRights` |
| **In which language is the document written?** | `dcterms:language` |
| **Where can I consult the document?** | `dcterms:identifier` |

## 4. Legal Knowledge Graph Construction: LLM and NLP Pipelines

Recent LLM advances have transformed *Natural Language Processing* (NLP) by enabling ontology development, knowledge graph construction, and query answering. These models enhance accuracy and efficiency in information extraction, addressing limitations of traditional knowledge engineering methods [21]. These models can augment traditional approaches by leveraging extensive training data to infer relationships, identify patterns, and construct meaningful representations of unstructured textual data, such as court rulings and legal documents.

Legal document analysis technology based on LLMs, with the semantic understanding capacity and contextual modeling ability, may effectively tackle these challenges and provide powerful auxiliary tools for [22]. This synthesis of LLM and KG enables better understanding and generation of structured information, addressing long-standing challenges in the field. Developing legal KGs is challenging due to the domain complexity. This approach combines LLMs for ontology generation and data extraction with traditional NLP for structured, precise text processing. Integrating both methods into a unified pipeline enables the creation of accurate, comprehensive KGs aligned with domain requirements and validated using CQs.

Effective *prompt engineering* [23] is essential for leveraging LLMs to produce desired outputs. This approach utilized two key techniques: *zero-shot prompting*, suitable for simple tasks with clearly expressed requirements, and *few-shot prompting*, which includes examples to guide the model in handling more complex tasks and reducing errors like hallucination.

### *4.1. KG Development*

The proposed methodology for generating KGs from legal rulings integrates advanced LLMs and traditional NLP techniques into a cohesive workflow. It involves two pipelines, one of which is an NLP-based pipeline[25]. The LLM pipeline includes document preparation, RAG (*Retrieval Augmented Generation*) development, ontology and KG creation, and competency question validation, while the NLP pipeline focuses on preprocessing, POS tagging, and triple extraction for constructing KGs.

[25] https://github.com/Fra3005/PreJust4Womans

#### 4.1.1. LLM's Pipeline

**Document Preparation:** The case study focused on analyzing and constructing KG from legal judgments, considering two distinct types of input. The first type utilized the entire judgment (full-text), while the second involved a specific sub-part of the judgment, selected by domain experts. The aim was to assess the effectiveness of both input types. For preparation, automated functions were used to extract text from PDFs for the full-text input. However, due to the varying structure and complexity of each judgment, manual extraction was employed for the sub-part input. Once the text was extracted, it was formatted into a document, ready for the subsequent steps.

**RAG's Creation:** RAG models have been employed to limit the output of LLMs to the context of each specific document. Although LLMs are trained on vast amounts of data, sometimes they struggle with providing accurate responses to specialized tasks. To address this issue, RAGs enhance traditional LLMs by integrating well-defined corpora [24], enabling LLMs to access two types of data when responding to user queries: parametric data, which refers to the model's training data, and nonparametric data, which consists of new, external data stored in a vector container. The LLM inspects this nonparametric data to generate precise answers. For the creation of RAGs, the TogetherEmbedding method[26] from the LangChain library was used. BERT-M2 served as the embedding model for semantic search within the vector dataset, and FAISS[27], accessed through the langchain_community library, was used to manage the vector dataset. RAGs were created for all five documents involved, which were then used as the contextual basis for the LLM to generate responses to various prompts.

**Base's Ontology Creations:** At this stage, the ontology, serving as a T-box for constructing the final KGs, was developed. Initially, GPT-4.o was employed to generate a foundational ontology, as Mixtral 8x22b proved incapable of creating a general schema for this topic. Using GPT-4.o resulted in a basic ontology containing core elements such as classes (e.g., *Abuse and LegalCase*) and *ObjectProperty*. Subsequently, the ontology was enriched with additional elements using Mixtral, which, despite explicit instructions in the prompts, tended to add extra components like *DataProperty* and *ObjectProperty* during the KG generation process. To further refine the ontology, a zero-shot prompting technique was applied to domain-specific documents. Prompts instructed the model to expand the ontology by incorporating fundamental concepts from these documents. This process was repeated with several randomly selected documents and, after multiple iterations, it was observed that Mixtral consistently produced similar elements regardless of the input document. This indicated structural similarities among the documents, allowing the newly generated elements to be reused across them, effectively identifying common patterns. Once the ontology was finalized, a thorough review was conducted to eliminate superfluous elements in the schema. This step addressed inconsistencies and redundancies in *DataProperty* elements, ensuring the alignment with the project requirements.

**KG Creation:** Building on the created ontology and the RAG for each document, a prompt engineering phase was undertaken to generate KGs for each document. This phase utilized a few-shot prompting technique to design effective prompts. Once the individual KGs were generated, they were merged into a unified graph. The merging pro-

[26] https://python.langchain.com/v0.2/docs/integrations/text_embedding/together/
[27] https://python.langchain.com/v0.1/docs/integrations/vectorstores/faiss/

**Table 2.** Score CQs answering of Large Language Models Approach.

| Results | Full-text | Sub-part |
|---|---|---|
| **Which legal case is associated with Violation?** | 3 | 4 |
| **What is the legal outcome of Case?** | 4 | 4 |
| **What is the reason stated for the judgment?** | 4 | 2 |
| **What abuse is related to Judgment?** | 5 | 4 |
| **What is the severity level of Abuse?** | 2 | 2 |
| **What is the duration and frequency of Abuse?** | 3 | 2 |
| **Which legal articles are violated?** | 4 | 4 |
| **What is the context of Abuse?** | 5 | 4 |
| **Which court judged the Case?** | 2 | 1 |
| **What are the damages related to Judgment?** | 1 | 3 |
| **What are the consequences of Abuse?** | 5 | 4 |
| **How much in legal damages and costs were awarded?** | 0 | 4 |
| **What is the legal status?** | 5 | 3 |
| **Total** | **40/65** | **37/65** |

cess preserved the distinct entities from each document and ensured consistency across the combined KG.

**CQ's Creation and Answering:** CQs were generated with Mixtral based on the ontology, with a subset selected for use. Mixtral then produced zero-shot answers for these CQs, which were manually verified against the KG and source text to avoid hallucinations. The validated results were stored in a CSV file for further analysis.

#### *4.1.2. NLP's Pipeline*

**Preprocessing:** This step, as described above, was conducted to evaluate and compare new LLM-based approaches. The process was applied to only one of the five previously analyzed documents. Initially, after the document was read and its text extracted, a preprocessing phase was carried out using the NLTK library [28]. This phase involved several steps, including punctuation removal, tokenization, stopword removal, and lemmatization, to prepare the text for subsequent analysis.

**Part of Speech:** After tokenization, the SpaCy library was used to perform POS tagging for each token. POS tagging is a NLP technique that categorizes the components of a sentence, making it particularly valuable for complex tasks like Sentiment Analysis or Machine Translation. In this study, POS tagging was utilized to extract triples from the document. Tokens were categorized into "subj" (subject), "verb," and "obj" (object) based on their grammatical roles, enabling the creation of structured triples.

**Triples Creation:** Finally, triples were constructed by associating each identified subject with its corresponding verb and object. This process resulted in a structured representation of the text, capturing meaningful relations between entities within the document.

### *4.2. Empirical Evaluation and Results*

The LLM-based KG generation was qualitatively evaluated using CQs. Across 5 documents with 13 CQs each, full text inputs produced 61.5% consistent answers, while

[28] https://www.nltk.org/

sub-part inputs yielded 56.9%. This shows a trade-off: full-text offers broader coverage, whereas sub-parts provide more focused and relevant results, particularly for factual or procedural queries, offering insight into the KG structure, coverage, and correctness.

This approach assesses LLMs in knowledge engineering tasks such as ontology generation, instance creation, and comparison with traditional NLP methods. LLMs struggled to create complete domain-specific ontologies using only pretrained knowledge, requiring additional documents and expert oversight. While the generated ontologies generally performed well, instances were sometimes incomplete or inaccurate, highlighting the need for human validation. Two specific use cases were explored to evaluate the effectiveness of LLMs in document processing:

1) *Full-text approach:* This use case involved processing the entire document to construct KGs and answer CQs. While this approach provided the highest quantity of responses due to its comprehensive coverage, the relevance and accuracy of the responses varied. Full-text processing captured more nuanced information but required significant manual verification to filter out irrelevant or incorrect outputs. The technical challenges included longer processing times and token size limitations inherent in LLMs.

2) *Sub-part approach*: The approach used expert-selected text segments to focus on the most relevant content. While it generated fewer responses than using the full text, it was faster, avoided token limits, and better aligned with expert objectives, though it relied on careful expert input to avoid missing important information.

The evaluation of the CQs (Tab. 2) emphasized responses aligned with the text, excluding incomplete entities generated solely by the LLM. The resulting KG, available in the GitHub repository[29], structures knowledge from legal and policy documents on women's rights and gender-based issues. It defines 12 classes, 9 object properties, and 17 data properties, outlining the domain's conceptual structure without instances, and serves as a basis for instantiating real-world scenarios and answering CQs.

The CQs were validated for accuracy against source documents. Comparing the two use cases showed that the full-text approach yields more information, while the sub-part approach is better for targeted, domain-specific tasks. This highlights the need to balance data coverage with precision, guided by expert input, to optimize LLM results.

## 5. Comparative Overview: Bottom-Up vs LLM-Based Approaches

We have proposed two distinct methodologies for constructing KGs based on sentences on cases of violence against women: a bottom-up approach and an LLM-based approach. This section discusses the two approaches, comparing their methodologies, outputs, and suitability for our purpose. A summary of the discussion is provided in Tab. 3.

The bottom-up approach yields precise, domain-specific KGs aligned with resources like ECLI and Wikidata, supporting formal reasoning and SPARQL but demanding significant effort and offering limited adaptability. The LLM-based approach, using zero-/few-shot prompting and retrieval-augmented generation, enables scalable, flexible, and rapid KG creation, though validation is essential to mitigate inaccuracies. Together, they are complementary: bottom-up ensures semantic accuracy, while LLMs support fast prototyping and large-scale processing.

[29] https://github.com/Fra3005/PreJust4Womans/blob/main/Commons/final_onto.ttl

**Table 3.** Comparative Analysis of Bottom-Up and LLM-Based Approaches.

| Criteria | Bottom-Up Approach | LLM-Based Approach |
|---|---|---|
| **Methodology** | Structured pipeline | Combine LLM capabilities with NLP and RAG techniques |
| **Strengths** | High precision and semantic alignment; Reliable for domain-specific; Structured tasks | Fast and scalable KG creation |
| **Challenges** | Labor-intensive and time-consuming; less adaptable to novel or unexpected patterns | Accuracy issues (hallucination and token limits |
| **Performance on CQs** | Excels in precise, structured queries using semantic standards | Generate broader responses but needs post-processing to ensure relevance |
| **Use Case Suitability** | Best for formal legal reasoning and SPARQL-based tasks | Ideal for exploratory tasks, large scale processing, and rapid prototyping |
| **Scalability** | Limited scalability due to manual data processing | Highly scalable for diverse and large datasets |
| **Output Consistency** | Outputs are precise and domain-aligned | Outputs may vary in consistency and require refinement |
| **Ontology Adaptability** | Limited to predefined ontologies | Highly adaptable to diverse and nuanced data patterns |
| **Ontology Concepts** | Higher initially, with many specific concepts emerging from data | More limited, well-organized in a clear hierarchy |
| **Ontology Relations** | More dynamic, based on frequently occurring relations in data | Accurate and based on logical modeling |
| **Ontology Size** | Large in size (583.7 KB), including a more extensive set of classes, properties, and relationships | Smaller in size (6.4 KB), suggesting a more concise ontology with potentially fewer classes and relationships |

**Ontology Development:** The project produced two ontologies: one generated with an LLM, which was efficient but generic, and another developed bottom-up, which was semantically richer and aligned with legal reasoning. Their comparison shows the complementary strengths of automation and expert-driven design, emphasizing the need for refinement to ensure legal fidelity. This hybrid approach demonstrates the feasibility of combining structured extraction with LLM techniques for scalable KG generation, supporting broader access to legal data and enhancing interpretability of predictions.

**Ensuring FAIRness in Legal Knowledge Engineering:** The LKG complies with the FAIR *(Findable, Accessible, Interoperable, Reusable)* principles. It is *findable* via a persistent DOI[30] and its registration in the Linked Data Cloud[31]. It is *accessible* through Zenodo, GitHub, and a public SPARQL endpoint. The KG ensures *interoperability* by using the RDF/Turtle format and standard vocabularies, and is *reusable* under a CC BY 4.0 license with full documentation and source code on GitHub.

## 6. Conclusion

With this paper, we introduced a novel Legal Knowledge Graph focused on sentences about cases of violence against women. Designed in accordance with FAIR principles, the Legal KG ensures interoperability and possibly reusability across legal and AI systems. To construct the KG, two automated methodologies have been adopted/tailored: a bottom-up approach and an LLM-based approach. Each methodology offers distinct advantages: precision and domain-specific accuracy for the case of the bottom-up solution; scalability and flexibility for the case of the LLM-based approach.

Future work will focus on merging the two KGs into a unified resource, to be possibly expanded, and on extending the KG applicability to the automated detection of legal patterns. Additionally, we are planning to experiment the applicability of the proposed methodologies to other legal domains in order to showcase the generality of the proposed solutions and ultimately contributing to a unified, FAIR-aligned representation of European laws that may enhance legal automation and judicial efficiency.

[30] https://doi.org/10.5281/zenodo.15270173

[31] https://lod-cloud.net/dataset/PREJUST4WOMAN_PROJECT

*Acknowledgments*

This work was partially supported by project *FAIR - Future AI Research* (PE00000013), spoke 6 - Symbiotic AI (https://future-ai-research.it/) under the PNRR MUR program funded by the European Union - NextGenerationEU, and by PRIN project *HypeKG - Hybrid Prediction and Explanation with Knowledge Graphs* (Prot. 2022Y34XNM, CUP H53D23003700006) under the PNRR MUR program funded by the European Union - NextGenerationEU.

## References

[1] Porto FD, Fantozzi P, Naldi M, Rangone N. Mining EU consultations through AI. Artificial Intelligence and Law Article. 2024.

[2] Li J, Qian L, Liu P, Liu T. Construction of Legal Knowledge Graph Based on Knowledge-Enhanced Large Language Models. Information. 2024;15(11):666.

[3] Westermann H, Savelka J, Walker VR, Ashley KD, Benyekhlef K. Computer-Assisted Creation of Boolean Search Rules for Text Classification in the Legal Domain. In: Proceedings of JURIX 2019; 2019. .

[4] Hogan A, Gutierrez C, Cochez M, de Melo G, Kirrane S, Polleres A, et al. Knowledge Graphs. Synthesis lectures on data, semantics, and knowledge. Morgan & Claypool Publishers; 2022.

[5] Wang C, Su W, Hu Y, Ai Q, Wu Y, Luo C, et al. LeKUBE: A Legal Knowledge Update BEnchmark. arXivorg. 2024;abs/2407.14192.

[6] Wang X, Zhang X, Hoo V, Shao Z, Zhang X. LegalReasoner: A Multi-Stage Framework for Legal Judgment Prediction via Large Language Models and Knowledge Integration. IEEE Access. 2024:1-1.

[7] Dawa I, Genene M. Violence against Women. In: Kurtz LR, editor. Encyclopedia of Violence, Peace, & Conflict. 3rd ed. Elsevier; 2022. .

[8] Tamašauskaite G, Groth P. Defining a Knowledge Graph Development Process Through a Systematic Review. ACM Transactions on Software Engineering and Methodology. 2022;32:1 40.

[9] Li J, Garijo D, Poveda-Villalón M. Large Language Models for Ontology Engineering: A Systematic Literature Review. SWJ. 2025. Under review.

[10] Monfardini GKQ, Salamon JS, Barcellos MP. Use of Competency Questions in Ontology Engineering: A Survey. In: Almeida JPA, Borbinha J, Guizzardi G, Link S, Zdravkovic J, editors. Conceptual Modeling. Cham: Springer Nature Switzerland; 2023. p. 45-64.

[11] Zhao B, Zhao Y, Mao Y. A Method for Judicial Case Knowledge Graph Construction Based on Event Extraction. In: Proceedings of the 2024 9th International Conference on Intelligent Information Technology; 2024. p. 62-9.

[12] Stavropoulou S, Romas I, Tsekeridou S, Loutsaris MA, Lampoltshammer T, Thurnay L, et al. Architecting an innovative big open legal data analytics, search and retrieval platform. In: Proceedings of the 13th international conference on theory and practice of electronic governance; 2020. p. 723-30.

[13] Wei H. Intelligent Legal Document Generation System and Method Based on Knowledge Graph. In: Proceedings of the 2024 International Conference on Machine Intelligence and Digital Applications; 2024. p. 350-4.

[14] Francesconi E, Governatori G. Patterns for legal compliance checking in a decidable framework of linked open data. Artificial Intelligence and Law. 2023;31(3):445-64.

[15] Oliveira Fd, Oliveira JMPd. A RDF-based graph to representing and searching parts of legal documents. Artificial Intelligence and Law. 2024;32(3):667-95.

[16] Schneider JM, Rehm G, Montiel-Ponsoda E, Rodríguez-Doncel V, Martín-Chozas P, Navas-Loro M, et al. Lynx: A knowledge-based AI service platform for content processing, enrichment and analysis for the legal domain. Information Systems. 2022;106:101966.

[17] Anelli VW, Brienza E, Recupero M, Greco F, De Maria A, Noia D, et al. Navigating the legal landscape: Developing Italy's official legal knowledge graph for enhanced legislative and public services. In: Falchi F, et al., editors. Proc. of Ital-IA 2023. vol. 3486. CEUR; 2023. p. 223-8.

[18] Vuong THY, Hoang MQ, Nguyen TM, Nguyen HT, Nguyen HT. Constructing a Knowledge Graph for Vietnamese Legal Cases with Heterogeneous Graphs. In: Proc. of IEEE-KSE 2023. IEEE; 2023. .

[19] Sovrano F, Palmirani M, Vitali F. Legal Knowledge Extraction for Knowledge Graph Based Question-Answering. In: Proc. of JURIX 2020. Frontiers in Artificial Intelligence and Applications. IOS; 2020. p. 143-53.
[20] Filtz E, Kirrane S, Polleres A. The linked legal data landscape: linking legal data across different countries. Artificial Intelligence and Law. 2021.
[21] Chen L, Xu J, Wu T, Liu J. Information Extraction of Aviation Accident Causation Knowledge Graph: An LLM-Based Approach. Electronics. 2024;13(19):3936.
[22] Ammar A, Koubaa A, Benjdira B, Nacar O, Sibaee S. Prediction of Arabic Legal Rulings Using Large Language Models. Electronics. 2024;13(4):764.
[23] Ye Q, Axmed M, Pryzant R, Khani F. Prompt Engineering a Prompt Engineer; 2024. Available from: https://arxiv.org/abs/2311.05661.
[24] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems. 2020;33:9459-74.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*

*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251576*

# Explaining Discretionary Judicial Decisions Using Answer Set Programming

Josephine DIK
*TU Wien*

**Abstract.** We introduce a method for analyzing discretionary judicial decisions by testing the internal consistency of judges' reasoning and uncovering their underlying assumptions. The approach represents both the facts of a case and the judge's stated evaluation in Answer Set Programming (ASP) and uses a Python analysis to examine the resulting answer sets. Applied to a child custody case, the method reveals the weights implicitly assigned to different factors, highlights when the reasoning cannot support the outcome, and identifies alternative weightings that would make the decision consistent. The contribution is a formal tool for making judicial reasoning more transparent, analyzable, and open to critique.

**Keywords.** explanations, answer set programming, judicial discretion, judicial decision-making, deontic concepts

## 1. Introduction & Approach

Judicial discretion is an inescapable and integral part of most legal systems. It requires the judges to weigh competing considerations, interpret norms, and make decisions where the law leaves room for subjective judgment. However, discretion does not imply arbitrariness, and the more a decision relies on discretionary balancing, the more its legitimacy depends on a transparent and logically coherent justification [1,2].

Despite the central role of judicial discretion in legal reasoning, it has received limited attention in the field of AI&Law. Most computational models of legal reasoning emphasize rule formalization or case-based inference, often treating decisions as following deterministically from stated inputs. Traditionally, discretion is considered to be one of the main limitations of such a derivation mode [3,4,5,6,7] and is therefore rarely addressed in logic-based methods.

Our earlier work begins to address this gap by formalizing the normative characteristics of discretionary decision-making in child custody cases [8,9]. In the current paper, we take the next step: rather than modeling discretion itself, we test whether the justification the judge gives for their discretionary decision is logically consistent, and we expose their underlying assumptions.

To do so, we use Answer Set Programming (ASP) [10]: a declarative framework and one of the most well-known and successful paradigms for knowledge representation and reasoning. We use ASP to generate all *consistent* explanations of judicial decisions given a certain case, and use Python to examine the conclusions that can be drawn about the reasoning of the judge. Rather than replacing judicial reasoning, the tool aims to reconstruct the implicit weightings and trade-offs that explain a given decision.

We apply our framework to an area of law where the use of discretion is apparent: child custody cases. Here, the only principle to serve is that the decision must be in the best interest of the child. Given the variety of features influencing what is in the best interest of a child in a given family (divorce) situation, the Hungarian[1] statutory law (both the Civil Code and the Family Act) leaves it to the judge to deliberate these features. The Supreme Court's Directive nr 17[2] provides some further details pointing out that the judge must carefully examine whether the parents' personalities, lifestyles, and moral qualities make them suitable for raising the child. Moreover, Directive 17 provides framing regarding how *free* the judge is in their deliberation. It is pointed out that giving excessive weight to a particular circumstance and ignoring other aspects goes against properly enforcing the child's best interest when deciding custody. This set of provisions affirms that discretion is allowed in the assessment, particularly in assigning the importance and weight of each factor.

In our implementation, we build on the framework presented in [9], where the normative characteristics of judicial discretion in child custody cases are captured using ASP. We make use of their encoding of the case structure. Here, each parent is associated with a set of properties that have a normative status. A property is *obligatory* (e.g., providing safety) when it is obligatory for a judge to make a decision for a parent with that property. Conversely, a *prohibition* for a property (e.g., being in jail) implies that it is prohibited for a judge to make a decision for a parent with that property. Thus, to be eligible for custody, a parent must have all obligatory properties and none of the forbidden ones. Once both parents are eligible, the remaining properties are interpreted as three types of *nuanced permissions*: 1. *positive permissions*, making a decision that is not only permitted but even encouraged, 2. *negative permissions*, being discouraged, while still allowed, 3. *neutral permissions*, where no preference is taken. The nuances allow us to describe the effect of deliberation, differentiating between making a decision that is not only permitted but even encouraged, from those being discouraged, while still fine, or those being neutral. These permitted properties are *weighted* according to their importance for the child's development using natural numbers, and these result in the final decision. In this work, we merely focus on the *permitted* properties.

The final procedure is as follows. First, we input a written case as well as a justification of a judge into a GPT-4o component. This component formalizes the case into an ASP encoding. This encoding consists of the properties of each parent, whether these properties are *good* or *bad*, whether the judge considers these properties *very important*, *important*, or *not so important*, and the decision of the judge for either parent A or parent B. Then, the ASP solver checks whether there are any models satisfying this encoding, and if so, it looks for all possible weight assignments for these properties. We refer to the three labels very important, important, and not so important as the *importance categories*. If a judicial justification is consistent, we employ Python to systematically examine the structure of all possible weight assignments generated by the ASP encoding. In this analysis, we extract the predicates that are present in all answer sets, which helps identify the assumptions that are implicitly necessary to explain a decision, and analyze relations between the weights of a property across all answer sets. We also compute the

[1] To stay consistent, we have looked into one national legal system, but it could have been any one of them, we suppose that the main setup and considerations are similar (at least in the European countries).

[2] https://kuria-birosag.hu/hu/iranyelvek/17-szamu-iranyelv Strictly speaking this Directive is not applicable anymore, but judges tend to act still in its spirit hence its content still prevails in the Hungarian judicature.

minimum and maximum weights and differences between property weights across all models. Moreover, we examine subsets of answer sets grouped by specific orderings of properties. We use GPT-4o to generate the ASP encoding, as this shows the potential for automating the entire process in the future, where a written case is the input, and the final analysis is the output.

While the weights are represented by natural numbers, we acknowledge that judges do not reason in *explicit* numeric terms. Their decisions, however, imply a relative importance that simple importance categories (or other labels) cannot fully capture. For example, assume four properties $p,q,r,s$ are labeled *very important*, but ordered as $p < q < r < s$. If Parent A has properties $s$ and $p$ and Parent B has properties $q$ and $r$, the label alone does not determine the outcome. Numeric values make such distinctions explicit. Exactly because these weights are not explicit, our approach provides a formal means to reconstruct and analyze the assumptions that underlie a given decision, making it easier to identify inconsistencies, justify appeals, or subject decisions to critique. However, to mitigate the risks of relying solely on numeric weights, we use an intermediate layer of abstraction via the importance categories, and use our tool to extract more information on the *exact importance*. We select 21 as the maximum weight for this implementation for several reasons. First, it allows for a division into three equal-width categories, enabling fine-grained distinctions while maintaining interpretability. Second, our running example includes six properties, and this range supports sufficient expressiveness to represent diverse judicial priorities without overwhelming the ASP solver with complexity. While other upper bounds could have been chosen, 21 strikes a balance between legal granularity and computational tractability. Future work will explore whether alternate scales affect the stability or fairness of outcomes.

The remainder of this paper is structured as follows. Sec. 2 introduces the basics of ASP and its key features. Sec. 3 describes our encoding in ASP and the analysis with Python. Sec. 4 puts the framework into practice and applies it to a child custody case. Sec. 5 discusses related work and concludes the paper.

## 2. Answer Set Programming in a Nutshell

We provide the foundational background necessary to understand the formalization in ASP and present how we utilize ASP's key features. The concepts used are aligned with [10], which we also recommend consulting for those not familiar with ASP.

**Logic Programs and Rules.** ASP is a declarative programming paradigm that uses rules to find models (answer sets) and operates over a set of atoms $At$. A *literal* is either an atom $p \in At$ or its explicit negation $-p$. An *extended logic programme* [11] consists of rules of the form: $H_1;\ldots;H_m$ :- $B_1,\ldots,B_n,\text{not}\,B_{n+1},\ldots,\text{not}\,B_k$ where $m,n,k \geq 0$, and each $H_i$ and each $B_j$ are literals. Here, *not* denotes weak (default) negation. The comma on the right-hand side represents a conjunction, and the semi-colon a the left-hand side a disjunction. In this syntax, the *head* of the rule is the atoms on the left-hand side, $H_1,\ldots,H_m$, and the *body* those on the right-hand side is $B_1,\ldots,B_n$ and $\text{not}\,B_{n+1},\ldots,\text{not}\,B_k$. Intuitively, the rule states: if all $B_1,\ldots,B_n$ are true, and none of $B_{n+1},\ldots,B_k$ are true, then one of $H_1,\ldots,H_m$ must be true. When the head of a rule is empty, the body must not hold. For instance, the rule :- $p(1)$ states that $p(1)$ is false, eliminating any answer set containing $p(1)$.

The *reduct* of a programme $\Pi$ w.r.t. an interpretation $T$, denoted $\Pi^T$, is obtained by (1) removing rules with a default literal *not* $L$ in the body if $L \in T$, and (2) removing default literals from remaining rules. An interpretation $T$ is an *answer set* of $\Pi$ if it is minimal (under set inclusion) among all interpretations satisfying $\Pi^T$.

**Aggregates.** ASP allows for the direct use of aggregates, which are elaborately described in Chapter 7 of [10]. Specifically, the aggregate `#sum` is useful for our application, as it allows us to sum all values of the properties of a parent. This results in a final score for each parent, determining the choice of the judge.

**Multiple Answer Sets.** The output of an ASP programme is *all* minimal answer sets that are consistent with the rules given by the programme. In our approach, each answer set includes a possible numeric weight for each property, and thus each answer set represents a possible explanation of the decision of the judge.

**Non-monotonicity.** ASP follows a non-monotonic approach to reasoning, allowing the representation of incomplete information, see e.g. [12]. This allows conclusions to change as new information is added. For example, a choice to grant custody to a mother can result in a choice for the father when new properties are added to the case. Specifically, previous explanations are dismissed when more predicates are added to the case.

**Use of Python** Our approach builds on the multiple answer sets that an ASP programme outputs. In some cases, the number of answer sets can be very high, which requires a script to analyze the output. We do this using Python, described in Sec. 3.

**Remark 1** *The full code used to produce the results in this paper can be found on the following github repository: https://github.com/josephinedik/JURIX2025.git. All results were generated using* `clingo` *version 5.4.0 (see [13]) and Python version 3.6.7.*

## 3. Implementation

Our implementation can be found in Table 1. The predicates used are the following:

- *props(X)* denotes that $X$ is a property;
- *parent(P)* denotes that $P$ is a parent;
- *hasProp(P,X)* denotes that parent $P$ has property $X$;
- *obl(X)* resp. *perm(X)* denotes that a property $X$ is obligatory resp. permitted;
- *goodness(X)* resp. *badness(X)* denotes that property $X$ is good resp. bad;
- *permPlus(X)* resp. *permMin(X)* resp. *permZero(X)* denotes that $X$ is a positively resp. negatively resp. neutrally permitted property;
- *num(N)* denotes that $N$ is a natural number;
- *weight(X,N)* denotes that $X$ has a weight of $N$ where $N$ is a natural number;
- *values(X,N)* denotes that $X$ has a value of $N$ where $N$ is an integer;
- *choose(P)* denotes that parent $P$ was chosen by the judge;
- *points(P,N)* denotes that parent $P$ has $N$ points;
- *category(C)* denotes that $C$ is a category;
- *catProp(X,C)* denotes that a property $X$ has category $C$.

As usual, capital letters represent variables; we will use (words starting with) lower-case letters for constants. Rules (1)-(16) are directly imported from [9]. The rules depicted

| | |
|---|---|
| obl(-X) :- -perm(X). -perm(-X) :- obl(X). perm(-X) :- -obl(X). -obl(-X) :- perm(X). | (1) |
| :- obl(X), obl(-X). perm(X) :- obl(X). | (2) |
| -goodness(X) :- badness(X). -badness(X) :- goodness(X). | (3) |
| permPlus(X) :- goodness(X), props(X), perm(X). | (4) |
| permMin(X) :- badness(X), props(X), perm(X). | (5) |
| permZero(X) :- not badness(X), not goodness(X), props(X), perm(X). | (6) |
| num(1..21). | (7) |
| 1{weight(X,N): num(N)}1:- props(X). | (8) |
| weight(X,0) :- permZero(X). | (9) |
| values(X,0) :- weight(X,0). | (10) |
| values(X,Y) :- weight(X,Y), permPlus(X). | (11) |
| values(X,-Y) :- weight(X,Y), permMin(X). | (12) |
| points(P,X) :- X=#sum{N,Y :values(Y,N), hasProp(P,Y)}, parent(P). | (13) |
| choose(P) :- X>Y, points(P,X), points(Q,Y), parent(P), parent(Q). | (14) |
| :- X=Y, points(P,X), points(Q,Y), P!=Q. | (15) |
| :- choose(P), choose(Q), P!= Q. | (16) |
| category(unimp). category(imp). category(veryimp). | (17) |
| :- props(X), catProp(X,unimp), weight(X,N), N>7, num(N). | (18) |
| :- props(X), catProp(X,imp), weight(X,N), N<8, num(N). | (19) |
| :- props(X), catProp(X,imp), weight(X,N), N>14, num(N). | (20) |
| :- props(X), catProp(X,veryimp), weight(X,N), N<15, num(N). | (21) |
| 1{catProp(X,Y):category(Y)}1 :- props(X). | (22) |
| perm(X) :- props(X). | (23) |

**Table 1.** Core Code

on rows (1) and (2) represent the properties of the deontic operators. The rules on (1) state that obligation and permission are interdefinable; and (2) states that a property can not be both obligatory and prohibited, as well as that obligatory properties are permitted properties[3]. The rules on row (3) state that if a property is bad, it is not good, and vice versa. Rules (4)-(6) define the different types of permissions, where (4) is the positive permission, (5) the negative permission and (6) the neutral permission. Regarding the weights, we only consider natural numbers between 1 and 21 (7), and each property has exactly one weight (8). The final value of a property follows from the type of permission that holds and the weight (9)-(12): when a property is neutrally permitted, then it gets a weight of zero (9) and a value of 0 (10); if a property is positively permitted the value is the same as the weight; lastly, if a property is negatively permitted then the value is the weight negative (12). The points of a parent are calculated by summing all values of the properties that the parent has (13). The person with the most points is chosen (14). Since a judge has to make a decision, two parents can not have an equal amount of points (15) and only one parent can be chosen (16).

Outside of weights, properties are also divided into one of the three importance categories: not so important (*unimp*), important (*imp*), and very important (*veryimp*) (17). Rules (18)-(21) define how the properties are classified into categories: when the weight is 7 or lower, a property is not so important (18); if their weight is between 8 and 14, a property is important (19),(20); when the weight is 15 or higher, the property

[3]While we do not consider obligatory properties in this paper, we include these rules for completeness.

is important (21). Each property has to be put in exactly one category (22). As we only focus on permitted properties, we add rule (23).

*Analysis using Python*

A case description consists of a set of rules of the form *props(X)*, *parent(P)*, *hasProp(P,X)*, *goodness(X)*, and *badness(X)*. The justification of the judge is a set of rules of the form *catProp(P,C)*. To analyze all possible explanations given these rules, we developed a Python-based tool, which works as follows.

**Formalizing case using GPT-4o** As a prompt, we give the case description and the template of the ASP formalization, describing what the predicates mean. The exact prompts can be found here.

**Extracting Answer Sets** Using Clingo's Python API, we load the ASP programme and extract *all* answer sets. Each model is represented as a Python dictionary mapping property names to their weights, as our focus is on analyzing the weights.

**Global and Conditional Weight Relations** The main goal of the analysis is to identify generalizable numerical constraints that describe how weights relate across models. For each pair of properties $p$ and $q$, we search for minimal functions, namely: 1. multiplicative bounds and 2. difference bounds. *Multiplicative bounds* are functions of the form $p \geq k \cdot q$, which means that the weight of property $p$ is at least $k$ times as much as the weight of $q$, where $k$ is the largest natural number for which the inequality holds across *all* answer sets. These functions are first derived across all answer sets, then also conditionally within subsets that share a certain structure. Namely, for each property $p$ and $q$ for which the ordering is unknown, the tool analyzes the multiplicative bounds of the answer sets where $p$ is more important than $q$ and those where $q$ is more important than $p$. The ability to partition answer sets by property orderings and examine weight constraints within each subset is a unique advantage of ASP-based reasoning. It allows us to simulate multiple plausible judicial perspectives and identify where and how they diverge. Then, *difference bounds* are functions of the form $m \geq |p - q| \geq n$ where $m$ is the maximum difference of weight between two properties $p$ and $q$, and $n$ the minimum.

**Output** When the ASP encoding is *satisfiable*, the script prints the multiplicative bounds and the difference bounds. When the encoding is *unsatisfiable*, and there are no answer sets, we conclude the justification for the decision given by the judge is inconsistent. In this case, the script removes one by one the predicates of the form $catProp(X,C)$ to see which predicate causes the inconsistency. This allows us to find which property or which set of properties has a different weight than claimed by the judge.

## 4. Framework into Practice

We demonstrate the practical utility of our framework by applying it to a real-world Hungarian child custody case from [8]. Our goal is twofold. First, to test whether a judge's justification for a decision is logically consistent, and, if not, what minimal changes are needed to recover consistency. Second, to reveal implicit assumptions present in the justification of the judge w.r.t. the weight of the properties.

**Example 1 (Case Description)** *In this case, both parents are capable of raising children appropriately. Where differences can be found are the following: Parent B does not have a good relationship with the child (actually hit him once in a conflict). The child has a good relationship with Parent A; additionally, the child declared their wish to live with this parent. Parent A, when the deterioration of the marriage seemed to be final, decided to move first to Austria, then to Germany with the child without the consent of Parent B.*

Using GPT-4o, this case is encoded using the following predicates:

```
parent(a;b).
props(capable_of_raising_children).
props(good_relationship_with_child; child_wants_to_live_with_parent).
props(moved_without_consent; bad_relationship_with_child;hit_child_once).
hasProp(a, capable_of_raising_children; a, good_relationship_with_child).
hasProp(a, child_wants_to_live_with_parent; a, moved_without_consent).
hasProp(b, capable_of_raising_children; b, bad_relationship_with_child; b, hit_child_once).
goodness(capable_of_raising_children; good_relationship_with_child; child_wants_to_live_with_parent).
badness(moved_without_consent;bad_relationship_with_child; hit_child_once).
```

The judge granted custody to Parent B and provided the following justification.

**Example 2 (Justification Judge)** *Even though it is very bad that Parent B hit the child, the judge at the court of the first instance found that moving abroad was not only illegal but also exceptionally estranging the child from Parent B. While the child declared his wish to live with Parent A, the judge found that this is due to this conscious estrangement, hence not to be taken seriously and thus granted full custody to Parent B.*

The justification offered by the judge is encoded in ASP as follows:

```
catProp(capable_of_raising_children, unimp).
catProp(good_relationship_with_child, imp).
catProp(bad_relationship_with_child, imp).
catProp(child_wants_to_live_with_parent, unimp).
catProp(moved_without_consent, veryimp).
catProp(hit_child_once, veryimp).
choose(b).
```

Running our ASP model with the given justification resulted in *no answer sets*. This indicates that the justification is logically inconsistent with the final decision. The inconsistency was traced to the following lines.

```
1 Found 0 models.
2 Constraints at line 31 in 'file.lp' are causing unsatisfiability.
3 Constraints at line 29, 30 in 'file.lp' are causing unsatisfiability.
```

Line 31 corresponds *catProp(hit_child_once, veryimp)*, and the lines 29 and 30 correspond to *catProp(good_relationship_with_child,imp)* and *catProp(bad_relationship_with_child,imp)*, respectively. This means that the decision, as justified, cannot be supported under any consistent assignment of weights. Therefore, some properties were evaluated differently than claimed[4].

First, we evaluate the output without the constraint *catProp(hit_hild_once,veryimp)*. This reveals that all answer sets contain *catProp(hit_ child_once,unimp)*, and thus this

[4]We acknowledge there can be other reasons for this inconsistency, which we discuss in the conclusion.

property is deemed *unimportant*. Moreover, the maximum weight is 3 and the minimum weight 1. Conversely, the property *moved_without_consent* has a high weight, between 19 and 21.

From the justification of the judge, we assume they deem the hitting of the child still more important than the desires of the child, as the judge notes that the hitting of the child is very bad and that the desires should not be taken seriously. Therefore, we analyse the subset of answer sets where the property *hit_child_once* is more important than *child_wants_to_live_with_parent*, and look at the multiplicative bounds:

```
Analyzing hit_child_once > child_wants_to_live_with_parent:
bad_relationship_with_child ≥ 2·hit_child_once
bad_relationship_with_child ≥ 7·child_wants_to_live_with_parent
good_relationship_with_child ≥ 2·hit_child_once
good_relationship_with_child ≥ 7·child_wants_to_live_with_parent
moved_without_consent ≥ 19·child_wants_to_live_with_parent
```

This sheds light on *how* irrelevant the desires of the child are with respect to the other properties, and how big the difference between these weights is. As the exact value of these numbers depends on the maximum value given (in our case 21), we do not treat these as inherently meaningful. Instead, the numbers merely indicate whether a property *overwhelmingly* dominates another property or less so.

Second, we remove line 29 (*catProp(good_relationship_with_child, imp)*), and line 30 (*catProp(bad_relationship_with_child, imp)*), and add back line 31 (*catProp(hit_child_once, veryimp)*). Then, it shows that in all answer sets these properties are marked *unimportant*, with a maximum weight of 3 and a minimum weight of 1. Additionally, the weights of these properties are close to the weight of the property *child_wants_to_live_with_parent*, shown by the following output:

```
2≥|child_wants_to_live_with_parent−bad_relationship_with_child|≥0
2≥|child_wants_to_live_with_parent−good_relationship_with_child|≥0
```

Meaning that the desires of the child are approximately equally (un)important as whether the child has a good or bad relationship with the parent.

In both situations, whether we remove the importance of the relationship or the importance of hitting the child, both analyses reveal the following:

```
moved_without_consent ≥ 6·child_wants_to_live_with_parent
Max weight of moved_without_consent is: 21
Min weight of moved_without_consent is: 19
Max weight of child_wants_to_live_with_parent is: 3
Min weight of child_wants_to_live_with_parent is: 1
```

This judge deems indeed that moving without consent has a high weight, at least 19, while what the child wants is minimal, at most 3. An appeal judge looked at this case and disagreed with this assessment. Their reasoning is given in the following example:

**Example 3** *The appeal court reevaluated the situation and found that the first judge had given unjustifiably great weight to the move and the implied estrangement, not evaluating properly the opinion of the child himself, who is old enough to have a reliable opinion immune to significant manipulation, having an established point of view in which the bad behavior of Parent B rather played a role. Hence, the appeal judge granted full custody to Parent A.*

This was encoded as follows:

```
catProp(capable_of_raising_children, unimp; moved_without_consent, unimp).
catProp(good_relationship_with_child, imp;bad_relationship_with_child, imp).
catProp(hit_child_once, veryimp;child_wants_to_live_with_parent, veryimp).
choose(a).
```

This justification is logically consistent, resulting in 117649 models, where the weights of hitting the child and what the child desires lie between 15 and 21.

As shown in this section, we check for consistency, reveal what would need to change to recover consistency, and expose the underlying weight assumptions. Additionally, this case underscores the role of judicial discretion: while both judges work with the same facts, their differing value prioritizations lead to opposing outcomes. Our tool makes such discretion transparent and traceable, crucial for legal reasoning in systems striving for both fairness and accountability.

## 5. Related Work & Concluding Remarks

This paper introduced a novel approach for analyzing discretionary judicial decisions in child custody cases. By reverse-engineering judicial justifications through ASP models and systematically evaluating their consistency via Python analysis, we demonstrated that it is possible to uncover implicit assumptions and inconsistencies in judicial reasoning, even when decisions appear to be well-justified on the surface.

The AI&Law community has given little explicit attention to the development of formal methods for discretionary reasoning. Around the early 2000s, a line of work examined discretion directly in the context of prediction in Australian (family) case law using neural networks for knowledge discovery from database [14,15,16,17,18,19]. To our knowledge, our efforts in [8,9], where we introduced nuanced permissions[5] for the deliberation within judicial discretion, are the first to address the normative characteristics of judicial discretion and to approach judicial discretion using symbolic methods. First, using modal logic in [8] and later using ASP [9].

The use of ASP is justified due to its use of constraints and its clear connection to the theory of constraint satisfaction, see e.g. [26,27], which provides an established theoretical framework for analyzing the coherence and justification of legal reasoning. This theory defends principles of reasoning not on the basis of their validity but on their coherence with the inferential practice. Our notion of consistency aligns with the notion of coherence given here, where a justification of the judge is inconsistent when their assessment is incoherent with the final decision that has been taken.

ASP has been widely applied to legal reasoning. For example, [28] used deontic programming to formalize legal provisions in a system supporting judicial decision-making, with particular attention to burden of proof. Moreover, [29] developed an ASP-based knowledge system for representing and reasoning about legal norms. ASP's non-monotonic nature makes it well-suited for modeling deontic concepts and avoiding classical paradoxes. Examples include Deontic Equilibrium Logic with explicit negation

[5]Different types of preferenced permissions have their own history: Sanskrit Mīmāṃsā philosophers discussed better-not permissions, akin to our negative, permissions [20]. Suboptimal permissions and supererogatory actions were studied in [21,22], while neutral permissions align with liberties in legal theory [23,24,25].

[30], implemented in ASP in [31]; the use of weak constraints to resolve well-known deontic paradoxes [32]; handling compensatory violations [33]; and modeling conflicting obligations for "rightful machines" [34].

Legal explanations in ASP are an emerging area. Specifically, [35] used ASP for the explanation of a decision by a court, by giving the specific article in the law that indicated whether a person made a violation or not. Additionally, the work of [36] proposed a tool in criminal law where the rules are encoded in ASP, and non-deterministic rules are used to generate possible hypotheses for a decision. Then, [37] modeled "Discretionality to Act" in an ASP-based automated legal reasoning framework, demonstrating that ASP can handle discretionary contexts.

The tool presented in this paper still has various limitations, paving several paths to future work. Immediate future work will integrate all components into a single tool, where a written justification of a case is the input, this input is formalized into ASP by the GPT-4o component, the ASP component generates all answer sets, and the output is the analysis done by Python. We acknowledge that LLMs such as GPT-4o still need to be checked to ensure they create the correct ASP encoding; however, it shows promise for a fully automatizable process. Tools like these enhance legal reasoning, making it more transparent, analyzable, and subject to critique and therefore contribute to the right to explanation[6]. This aligns with the broader goals of AI&Law: not to replace judges, but to help legal systems become more consistent, transparent, and fair.

Remaining directions for future work are the following. First of all, we used numeric abstractions for weighting legally relevant properties. While we acknowledge that judges do not assign explicit numbers in practice, our framework treats these values as a mathematical metaphor for qualitative assessments. Nevertheless, further work is needed to assess how different encoding schemes (such as ordinal, fuzzy, or case-based weightings) might affect interpretability across legal domains. Due to our choice of weights, the search space for the ASP programme was $n^7$ where $n$ is the number of properties and 7 the number of weights a property can assume, given its category, but future work would include the discussion of possible constraints that reduce the number of generated explanations. In the running example, we assumed that the inconsistent justification of the judge was caused by a wrong categorization of the properties. However, there are alternative explanations, such as a property that has not been named in the judicial justification because the judge wanted to hide that they took this into account (such as race, gender, or any other discriminatory factors). Alternatively, a property could have been classified as negative by a judge, while it has a positive effect on the child's life. More violations are discussed in [9], and we explore these in extensions of this work. Our framework resembles that of [39], where case factors support or oppose conclusions, but we focus on justification consistency rather than value-based comparisons or precedent. Future work will extend our approach to case-based analysis [40].

While our example focuses on child custody, future research will generalize the framework to other discretionary legal domains.

[6] See, for example, [38] where a method involving abstract argumentation was proposed to meet this requirement for automated decision.

## Acknowledgements

This work was supported by the Marie Speyer Excellence Grant of 2024 *Formal Analysis of Discretionary Reasoning*, MSE-DISCREASON, (PI: Réka Markovich), also in part by the Austrian Science Fund (FWF - 6372-N) in the *Logical Methods for Deontic Explanations* project, and by the European Union's Horizon 2020 research and innovation programme under grant agreement No 101034440.  And we would like to thank Éva Püsök from the District Court of Esztergom for her valuable help in case selection.

## References

[1] Sartorius R. The justification of the judicial decision. Ethics. 1968;78(3):171-87.
[2] Pettys TE. Judicial Discretion in Constitutional Cases. JL & Pol. 2010;26:123.
[3] Raz J. The authority of law: essays on law and morality. New York: Oxford University Press; 1979.
[4] Shiner RA. Norm and nature: the movements of legal thought. New York: Oxford University Press.; 1992.
[5] Alexy R. A Theory of Legal Argumentation: The Theory of Rational Discourse as Theory of Legal Justification. Oxford University Press; 2009.
[6] Hart HLA. The Concept of Law. Raz J, Bulloch PA, editors. Oxford University Press; 2012.
[7] Dworkin R. Taking rights seriously. Bloomsbury Academic; 2013.
[8] Dik J, Markovich R. Modeling Judicial Discretion with Nuanced Permissions. In: JURIX. IOS Press BV; 2024. p. 48-59.
[9] Dik J, Markovich R. Judicial Discretion as Normative Reasoning: Deontic Characterization of Judicial Decision Making with Answer Set Programming. In: International Conference on AI and Law (ICAIL); 2025. p. 258-68.
[10] Lifschitz V. Answer Set Programming. Springer; 2019.
[11] Gelfond M, Lifschitz V. Classical negation in logic programs and disjunctive databases. New generation computing. 1991;9:365-85.
[12] Brewka G, Eiter T, Truszczyński M. Answer set programming at a glance. Communications of the ACM. 2011;54(12):92-103.
[13] Gebser M, Kaminski R, Kaufmann B, Ostrowski M, Schaub T, Wanko P. Theory solving made easy with clingo 5. In: Technical Communications of the 32nd International Conference on Logic Programming (ICLP 2016). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik; 2016. p. 2:1-2:15.
[14] Schild UJ, Zeleznikow J. A taxonomy for modelling discretionary decision making in the legal domain. In: Proceedings of the 10th international conference on Artificial intelligence and law; 2005. p. 60-4.
[15] Kannai R, Schild U, Zeleznikow J. Modeling the evolution of legal discretion. an artificial intelligence approach. Ratio Juris. 2007;20(4):530-58.
[16] Stranieri A, Zeleznikow J. The Role of Open Texture and Stare Decisis in Data Mining Discretion. In: Legal Knowledge Based System JURIX 1998; 1998. p. 100-11.
[17] Kerner YH, Schild U, Zeleznikow J. Developing computational models of discretion to build legal knowledge based systems. In: Proceedings of the 7th International Conference on Artificial Intelligence and Law. ICAIL '99. New York, NY, USA: Association for Computing Machinery; 1999. p. 206–213.
[18] Stranieri A, Yearwood J, Zeleznikow J. Tools for World Wide Web based legal decision support systems. In: Proceedings of ICAIL '01. New York, NY, USA: Association for Computing Machinery; 2001. p. 206–214.
[19] Stranieri A, Zeleznikow J, Gawler M, Lewis B. A hybrid rule–neural approach for the automation of legal reasoning in the discretionary domain of family law in Australia. Artificial Intelligence and Law. 1999;7:153–183.
[20] Ciabattoni A, Dik J, Freschi E. Disambiguating Permissions: A contribution from Mīmāmsā. In: Deontic Logic and Normative Systems 16th International Conference (DEON 2023); 2023. p. 99-118.
[21] McNamara P. Doing well enough: Toward a logic for common-sense morality. Studia Logica. 1996;57:167-92.
[22] McNamara P. Extensions and Variations of the DWE Framework with Applications. In: Deontic Logic and Normative Systems 16th International Conference, DEON 2023; 2023. p. 177-97.

[23] Kelsen H. Pure theory of law. Univ of California Press; 1967.
[24] Bentham J. The works. vol. 3. Tait; 1843.
[25] Hart HLA. Essays on Bentham. Clarendon; 1982.
[26] Thagard P, Verbeurgt K. Coherence as constraint satisfaction. Cognitive science. 1998;22(1):1-24.
[27] Araszkiewicz M. Limits of constraint satisfaction theory of coherence as a theory of (legal) reasoning. In: Coherence: Insights from philosophy, jurisprudence and artificial intelligence. Springer; 2013. p. 217-41.
[28] Beierle C, Freund B, Kern-Isberner G, Thimm M. Using defeasible logic programming for argumentation-based decision support in private law. In: Computational Models of Argument. IOS Press; 2010. p. 87-98.
[29] Aravanis T, Demiris K, Peppas P. Legal reasoning in answer set programming. In: 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI). IEEE; 2018. p. 302-6.
[30] Cabalar P, Ciabattoni A, van der Torre L. Deontic equilibrium logic with explicit negation. In: European Conference on Logics in Artificial Intelligence. Springer; 2023. p. 498-514.
[31] Cabalar P, Manteiga O. deolingo: Extending Answer Set Programming with Deontic Reasoning. Proceedings of the 17th International Conference on Deontic Logic and Normative Systems (DEON 2025):61-76.
[32] Hatschka C, Ciabattoni A, Eiter T. Deontic Paradoxes in ASP with Weak Constraints. arXiv preprint arXiv:230815870. 2023:367-80.
[33] Governatori G. An ASP implementation of defeasible deontic logic. KI-Künstliche Intelligenz. 2024;39:79-88.
[34] Wright AT. A Deontic Logic for Programming Rightful Machines: Kant's Normative Demand for Consistency in the Law. In: Bentzen B, Liao B, Xu T, editors. Logics for Ai and Law: Proceedings (Lingai) and (Lail); 2023. p. 208-22.
[35] Sileno G, et al. Answer Set Programming for Legal Decision Support and Explanation. In: Legal Knowledge and Information Systems: JURIX 2023: The Thirty-sixth Annual Conference, Maastricht, the Netherlands, 18-20 December 2023. vol. 379. SAGE Publications Limited; 2023. p. 383.
[36] Dovier A, Dreossi T, Formisano A, et al. XAI-LAW Towards a logic programming tool for taking and explaining legal decisions. In: CEUR WORKSHOP PROCEEDINGS. vol. 3733. CEUR-WS; 2024. .
[37] Arias J, Moreno-Rebato M, Rodriguez-García JA, Ossowski S. Automated legal reasoning with discretion to act using s (LAW). Artificial Intelligence and Law. 2024;32:1141-64.
[38] Liga LYD, Markovich R. Addressing the Right to Explanation and the Right to Challenge through Hybrid-AI: Symbolic Constraints over Large Language Models via Prompt Engineering. In: Proceedings of the 20th International Conference on AI and Law; 2025. p. 228-37.
[39] Bench-Capon T, Sartor G. A model of legal reasoning with cases incorporating theories and values. Artificial Intelligence. 2003;150(1-2):97-143.
[40] Canavotto I. Reasoning with inconsistent precedents. Artificial Intelligence and Law. 2025;33(1):137-66.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*

*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251577*

# Contrary-to-Duty Rights: From Hohfeld to Agreement Revision

Huimin DONG [a], Leendert VAN DER TORRE [b,c] and Liuwen YU [b]
[a] *TU Wien, Vienna, Austria*
[b] *University of Luxembourg, Esch-sur-Alzette, Luxembourg*
[c] *Zhejiang University, Hangzhou, China*

**Abstract.** We present a rights-first model of contrary-to-duty (CTD) reasoning with two remedial regimes and a revision track. In the *CTD-Claim* regime, when a primary duty is not fulfilled and no exception applies, a remedial claim detaches *automatically*, without any recognition act. In the *CTD-Power* regime, a remedial claim arises only if the rights-holder *exercises* a recognition power; until then there is no recognised violation and no remedial duty. *Revision-of-Duty* (RoD) is an alternative discretionary power that adapts the primary duty without recognising a violation, keeping the purpose aligned and avoiding sanctions. Under CTD-Power, *revision* competes directly with *recognition* on the same case. We develop a model for agreements, give concise dynamic-logic-style specifications of guards and acts, implement an institutional rights system that executes these specifications over live Hohfeldian bundles (with per-case exclusivity, exception handling and provenance), and show how agentic AI can use reasons to choose among the admissible acts within the rights-first framework.

**Keywords.** Hohfeldian Rights; Contrary-to-Duty reasoning; CTD-Claim; CTD-Power; Revision-of-Duty; Agreement Technologies; Normative Multi-Agent Systems

## 1. Introduction to the Rights-First Perspective

Contrary-to-duty (CTD) reasoning is usually presented in the context of *obligations*: if a primary obligation is not met, a secondary one applies [1, 2, 3, 4, 5]. Among Hohfeld's four correlatives of rights — *claim/duty*, *privilege (liberty)/no-claim*, *power/liability*, and *immunity/no-power* [6] — obligations correlate with only one correlative in legal settings, which is *claim/duty*. We study *contrary-to-duty (CTD) rights* as directed legal positions that become operative *because* some primary duty $\varphi$ was not fulfilled ($\neg\varphi$).

For example, let a lease require the tenant to pay rent on the first day of each month ($\varphi$). On the fifth day, payment has still not been made ($\neg\varphi$). This familiar situation are supported by the standard CTD forms:

- **CTD obligation (duty-first baseline).** "If rent is not paid on the first day of the month, the tenant must pay a late fee."
- **CTD permission (duty-first variant).** "If the landlord fails to provide heating, the tenant may withhold rent."

Now read the same situation through Hohfeld's four corners as *CTD-rights* that come into play *because* non-fulfillment $\neg\varphi$ occurs:

- **CTD-*Claim*.** On $\neg\varphi$ (no payment by the first day), and in the absence of exceptions, a *remedial claim* (e.g. late fee) *detaches automatically*; the landlord holds a claim, the tenant a correlative duty.
- **CTD-*Privilege* (Liberty).** On $\neg\varphi_L$ (e.g. no heating), the tenant holds a *privilege* to withhold rent; correlatively the landlord has *no-claim* to payment during the outage (a CTD permission as a Hohfeld legal relation).
- **CTD-*Power*.** On $\neg\varphi$, the lease gives the landlord a *recognition/enforcement power*: a remedial claim *arises only if* that power is exercised (there is no automatic duty otherwise).
- **CTD-*Immunity*.** On $\neg\varphi$ with a recognised *exception* (e.g. force majeure concerning payroll), the tenant holds an *immunity*: the landlord's recognition power is *disabled* in this case (no sanction can be validly created).

This example suggests that the traditional CTD obligation and permission are just two facets of a richer family of *CTD-rights*, and that their behaviours differ precisely along Hohfeld's correlatives.

Here, we study whether this plurality matters for agreements [7] in normative multi-agent systems (MAS) [8]. In our study, we observe that, after distinguishing the four CTD-rights, a crucial form of *power* comes into focus: *revision-of-duty* (RoD).

- **Revision-of-Duty (RoD)**. The parties *revise* the rent date (e.g. move it to the tenth day of this month). The purpose of the lease is preserved; the tenant is *immune* from violation-recognition with regard to the original date.

Figure 1 depicts our focus. CTD-CLAIM is situated at the *claim/duty* corner (a remedial claim detaches on $\neg\varphi$ unless an exception holds). CTD-POWER is situated at the *power/liability* corner (a remedial claim appears only when a recognition power is exercised). *RoD* is a *revision* power that changes the primary duty so that the same $\neg\varphi$ no longer counts as a violation and an *immunity* against recognition is created.

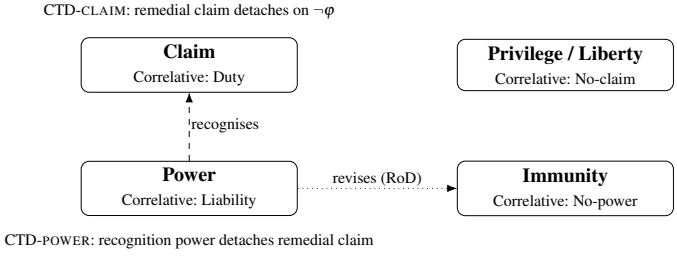



**Figure 1.** Hohfeldian rights with CTD markers. CTD-CLAIM: remedial claim detaches on non-fulfilment. CTD-POWER: remedial claim appears only if a recognition power is exercised. RoD is a *revision* power that adapts the duty and creates an immunity against recognition of the previous case.

The layout of this paper is as follows. Section 2 integrates our lifecycle intuition by applying CTD-CLAIM, CTD-POWER, and RoD to an agreement example. Section 3 explains how these ideas map to agreement states in an institution (purely at the level of concepts such as claim, power, and immunity). Section 4 briefly discusses what the conceptual picture implies for agentic AI & Law. Section 5 reviews related work, and Section 6 concludes the paper.

## 2. Specification of Agreements and Hohfedian Rights

We treat a signed agreement as a structured bundle of rights where the correlation between claim and duty is straightforward. To give another example, once the primary rights are activated, a buyer's claim to delivery induces a seller's duty to deliver, and breaches may trigger sanctions. Agreements are not static. Parties can exercise powers to either remedy a breach or revise the agreement in response to an unexpected situation by, for instance, making a delay permissible.

We adopt a *rights-first* perspective, reasoning about the CTD regime of secondaty rights in two distinct ways for the purpose of remedying issues in agreements. Building on the Hohfeldian rights with CTD markers in Figure 1, Figure 2 summarises the lifecycle we adopt of rights-based agreements: from *creation* to *primaries active*, an *issue* may trigger either the CTD track (with its two regimes) or the RoD track. A *revision* creates a new bundle and loops back to the active-bundle state, allowing *multiple revisions over time*. [1] This mirrors how agreements in practice may undergo several targeted amendments before final discharge, while keeping provenance auditable and conflicts controlled. The overall setting aligns with Agreement Technologies [7] and normative multi-agent systems (e.g. [8]), but uses Hohfeldian rights not merely for analysis but also to *operate* the state and evolution of agreements.

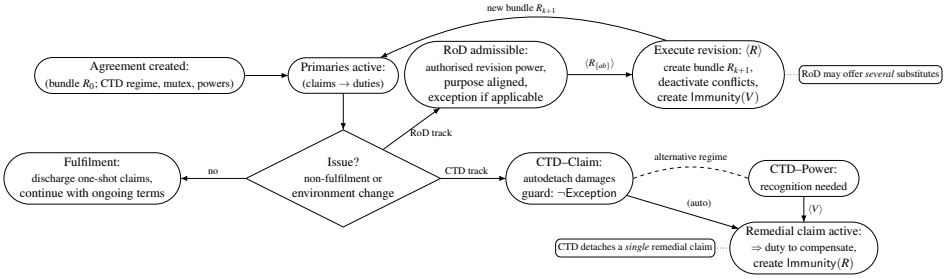



**Figure 2.** The agreement lifecycle with feedback. Its creation fixes issues with the regime and relevant powers. During execution, an issue triggers either CTD (automatic claim or recognition-based) or RoD (revision). CTD creates an immunity against revision; RoD creates an immunity against recognition and loops back to the *primaries active* state with the new bundle $R_{k+1}$, thereby allowing multiple revisions. Mutual-exclusion guards prevent conflicting content from co-existing; discharge/persistence govern how claims end.

*CTD-Claim Regime: Automatic Detachment.* In this first CTD-regime, secondary rights arise automatically. Suppose the seller fails to deliver on time without exception: the agreement stipulates that the buyer immediately holds a remedial claim (e.g. compensation). This is the CTD-Claim path. Under Hohfeld's framework, a remedial claim for the buyer entails a correlative duty for the seller. Importantly, no discretionary step is involved; the secondary duty detaches by operation of the law.

*CTD-Power Regime: Recognition-Based Detachment.* The second CTD-regime introduces discretion. Here, no remedial duty exists unless the rights-holder actively recognises a violation. In this CTD-Power regime, the buyer possesses a power to enforce, but until *exercised*, no claim arises. Recognition thus functions as a gatekeeping act: it transforms non-fulfilment into a violation and detaches the corresponding remedial duty.

[1] CTDs can repeat at multiple levels and may involve several layers of violation; see the detailed review in [9] for further discussion.

| Rights (Hohfeld bundles) | Claim | Power | Recognition | Revision | Programs (Power execution) |
|---|---|---|---|---|---|
| | $\mathsf{Claim}_{a\leftarrow b}(p/\neg q)$ | $\mathsf{Power}_{a\leftarrow b}(p/\neg q)$ | $\langle \mathsf{V}_b \rangle$ | $\langle \mathsf{R}_{\{ab\}} \rangle$ | |
| Language $\mathscr{L}$ is defined as $\varphi := p \in \mathit{Prop} \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid \mathsf{Claim}_{a\leftarrow b}(\varphi/\varphi) \mid \mathsf{Power}_{a\leftarrow b}(\varphi/\varphi) \mid \langle \mathsf{V}_b \rangle \varphi \mid \langle \mathsf{R}_{\{ab\}} \rangle \varphi$ | | | | | |

**Table 1.** Representation of Rights-First System = Hohfeld bundles Rights + Power Programs.

*Revision-of-Duty Track: Purpose-Preserving Adaptation.* Revision-of-Duty (RoD) is an alternative power that allows parties to adapt the agreement itself. Instead of treating non-fulfillment as a violation, the situation may be re-framed as an exception, prompting joint *revision*. For example, if delayed payment results from a temporary budget freeze, the parties may agree to extend the deadline. No breach is recognised, no sanction imposed, and the contract continues with a revised bundle of rights. RoD therefore embodies a *discretionary*, forward-looking power: it preserves contractual purpose and *immunises* the case from violation-recognition.

RoD is significant because it reveals a dimension absent from standard CTD analyses: *discretionary power*[2] to adapt duties. This distinguishes remedial enforcement from cooperative adjustment, and it highlights that the rights-first perspective is richer than duty-first models. Our specification is a *conceptual* contribution — it will discuss this conceptual distinction through a set of logical constraints while remaining intentionally minimal and without fully embedding formal semantics.

Seen through a Hohfeldian *rights-first* lens, this reasoning system provides a way to track whether situations are treated as *violations* or as *exceptions*. Accordingly, *agents* must decide whether to exercise their legal powers by recognising a violation or adopt a revision to cope with an exception.

The core elements of specifying rights-first reasoning are twofold: (i) rights as conditional norms [11], and (ii) programs that represent the exercise of those powers — such as violation recognition and agreement revision — as summarised in Table 1. Following Hohfeld's distinction between static and dynamic rights, we categorise rights into two main groups: claims and powers. Other legal rights such as privilege and immunity can be defined from these categories through Hohfeldian correlative relations.[3] The representation of programs extends this framework by capturing the idea of *exercising power*, thereby foregrounding the role of *agents' choices* within a dynamic logic formalisation [12].[4] For instance, $\langle \mathsf{V}_A \rangle \mathsf{Claim}_{M\leftarrow A}(\mathsf{Compensation}/\mathsf{NF})$ means that agent $A$'s exercise of violation-recognition results in a claim from $A$ against $M$ for compensation when the primary right is not fulfilled. By contrast, $\langle \mathsf{R}_{\{AM\}} \rangle \neg \mathsf{Claim}_{M\leftarrow A}(\mathsf{Compensation}/\mathsf{NF})$ means that agents $A$ and $M$'s joint revision leads to the negation of this claim-right. In this work, we focus on the specification; the semantics will be addressed in future work.

[2] Hart [10] maintained that the concept of discretion plays an important role in the legal domain. It involves making a choice that remains bound by legal values and principles, while keeping alternative options possible.

[3] Given the language $\mathscr{L}$, the propositional connectives $\vee, \rightarrow, \leftrightarrow$ can be defined in a standard way. By Hohfeld's system of correlatives, other fundamental legal positions can be defined as follows: $\mathsf{Duty}_{a\leftarrow b}(\varphi/\psi) := \mathsf{Claim}_{b\leftarrow a}(\varphi/\psi)$, $\mathsf{Liberty}_{a\leftarrow b}(\varphi/\psi) := \neg\mathsf{Claim}_{b\leftarrow a}(\neg\varphi/\psi)$, $\mathsf{Liability}_{a\leftarrow b}(\varphi/\psi) := \mathsf{Power}_{b\leftarrow a}(\varphi/\psi)$, and $\mathsf{Immunity}_{a\leftarrow b}(\varphi/\psi) := \neg\mathsf{Power}_{b\leftarrow a}(\varphi/\psi)$. For convenience, we write $\mathsf{Right}_{a\leftarrow b}(\varphi/\psi)$ to denote any element of this Hohfeldian bundle. The program operator $[\cdot]$ is the dual of $\langle\cdot\rangle$, e.g. $[\mathsf{R}_{\{ab\}}]\varphi := \neg\langle \mathsf{R}_{\{ab\}} \rangle \neg\varphi$.

[4] Here, we adopt the distinction between *legal competence* and *ontic ability* proposed by Markovich [13], and we focus on a lightweight representation of power programs as the exercise of legal competence. In this way, our specification remains as general as possible while still being able to capture the distinctions highlighted in existing frameworks.

*Example (Public Park Contract).* Suppose Municipality $M$ contracts Contractor $A$ to renovate a public park (Table 1).

**Primary right**: $A$ has the right to payment on time: $\mathsf{Right}_{M\leftarrow A}(\mathsf{Pay\text{-}timely}/\top)$.

If $M$ fails to pay on time (NF), two kinds of secondary rights arise: claims or powers.

**CTD-claim path:** Contractor $A$ may automatically obtain a claim-right against $M$ for compensation given the non-fulfillment — that is $\mathsf{Claim}_{M\leftarrow A}(\mathsf{Compensation}/\mathsf{NF})$.

Alternatively, the agreement may be structured so that Contractor $A$ holds a power — $\mathsf{Power}_{M\leftarrow A}(\mathsf{Compensation}/\mathsf{NF})$. This pathway leads to two different *agentic* tracks:

**CTD-power path:** $A$ has no immediate claim-right to compensation. The right materialises only if $A$ exercises its violation-recognition power, which then leads to a claim, as in the CTD-claim path: $\langle \mathsf{V}_A\rangle \mathsf{Claim}_{M\leftarrow A}(\mathsf{Compensation}/\mathsf{NF})$.

**RoD-power path:** Alternatively, $M$ and $A$ may classify the situation as an exception (EXP), e.g. they may extend the deadline because of a budgetary freeze. In this case, no violation is recognised, compensation is not triggered, and the contract adapts to preserve its purpose. *Exercising this revision power* [5] yields a new right: $\langle \mathsf{R}_{\{AM\}}\rangle \mathsf{Claim}_{M\leftarrow A}(\mathsf{30\text{-}days}/\mathsf{EXP})$.

To address conflicts, we introduce a lightweight representation of *consistency* principles for reasoning about Hohfeldian bundles, inspired by Hohfeld's (1913) theory [6].

- No contradictory claim-rights: $(C_1)$ $\neg(\mathsf{Claim}_{a\leftarrow b}(\varphi/\psi) \wedge \mathsf{Claim}_{a\leftarrow c}(\neg\varphi/\psi))$;
- Immunity as absence of power: $(C_2)$ $\langle \mathsf{V}_b\rangle[\mathsf{R}_{\{ab\}}]\bot$ and $\langle \mathsf{R}_{\{ab\}}\rangle[\mathsf{V}_b]\bot$.

Principle $(C_1)$ rules out conflict scenarios in which two contractors simultaneously demand mutually exclusive compensation from the Municipality when only one official contractor exists. Principle $(C_2)$ refines the notion of immunity in terms of the incompatibility of alternative actions: if a violation-recognition is acted upon, then no revision is executable, and vice versa — a form of *CTD–RoD program exclusivity*.

We impose three constraints that prevent conflicts among CTD-claims, CTD-powers, and RoD-powers and that ground our operational framework. $(P_1)$ *Auto-remedy excludes contrary powers*: if a CTD-Claim for $\varphi$ is active (automatic detachment), then no power is available that would, if exercised, create a claim for an incompatible $\psi$. $(P_2)$ *Mutually contrary CTD-powers do not co-exist*: two CTD-Powers whose exercise would yield mutually incompatible remedial claims must not be simultaneously available for the same case and locus (i.e. the same right-holder and counterparty). $(P_3)$ *Revision power may co-exist with a contrary CTD-power* (choice point): a RoD-power to create $\varphi$ may co-exist with a CTD- or RoD-power whose exercise would create a mutually incompatible $\psi$. Co-existence is allowed only at the level of availability; exercising one disables the other via track-specific immunities. Thus $P_1$ and $P_2$ enforce consistency, while $(P_3)$ preserves the discretionary choice that is central to the rights-first view.

We briefly comment on the specification of *exception* (EXP), the central notion in the rights system. If we wish to explicitly formalise the distinction between *exception* and *violation* — where exceptions (i) exclude sanctions and (ii) block recognition — two illustrative implications are: (1) $\mathsf{EXP} \rightarrow \langle \mathsf{R}_{\{ab\}}\rangle\neg\mathsf{Claim}_{a\leftarrow b}(\mathsf{sanction}/\psi)$ and (2) $\mathsf{EXP} \rightarrow \neg\langle \mathsf{V}_b\rangle\top$. That is, exceptions activate revision without sanction but never recognition.

[5] A revision power may be exercised individually or jointly, thereby detaching a new bundle of Hohfeldian rights. In contrast to existing frameworks [5, 14, 15], our lightweight specification provides a novel representation of rights detachment in multi-agent settings,

## 3. Institutional Rights System for Managing Agreements

The rights system we introduce in this section is a proof-of-concept institutional layer that maps observed *facts* and recorded *legal acts* to a live Hohfeld bundle for each agreement, while satisfying the constraints of the previous section.

Inputs are facts (performance logged, deadline passed), acts (RECOGNISE breach; authorised REVISE; ANNUL contract), and creation-time configuration (primary contents, CTD regime, admissible revisions, mutex constraints). Outputs are current **claims** (and induced **duties**), available **powers**, and any **immunities**, each with explicit *provenance* (*primary*, *CTD-claim*, *CTD-power+recognised*, *revision*). The system never chooses between admissible acts — that responsibility rests with the parties. By design, there is no duty without a live claim, every claim has schema or power provenance, CTD–RoD exclusivity applies on a per-case basis (via immunities), and exceptions disable recognition and create no sanctions. This stays faithful to the rights-first approach: *claims* and *powers* are primitive; *duties*, *liabilities*, and *immunities* arise as correlatives in states.

A useful analytical strategy is to separate *facts* from *acts*. Facts move the lifecycle without exercising power: *performance* discharges a one-shot normative element (e.g. primary, remedy or substitute); an *issue* (non-fulfilment) opens a case; an *exception* blocks recognition. Acts exercise powers and thus change bundles: RECOGNISE detaches the remedy (when available) and creates $\mathsf{Imm}(R)$; joint REVISE creates a substitute bundle and $\mathsf{Imm}(V)$; WAIT leaves the bundle unchanged; and ANNUL terminates the contract, leaving no rights remaining with the parties. This cleanly implements the conceptual contrast in Section 2 between *automatic* CTD-CLAIM and *discretionary* CTD-POWER.

Table 2 gives the state semantics.

- **S0 (Primaries)**: preconditions hold; primary claims induce duties; if configured, a revision power is available.
- **S1 (Discharge)**: a performed one-shot element — primary, remedy, or substitute — is removed; persistent terms continue.
- **S2 (Case open)**: non-fulfilment is recorded and the CTD regime is evaluated. In **CTD-Claim**, unless an exception holds, the remedy *detaches automatically* to **S3a**; there is no agent choice. In **CTD-Power**, the remedial posture depends on the *exercise* of that power; a **Choice** appears (**S3b**).
- **S3a (Remedy auto)**: remedy live and $\mathsf{Imm}(R)$ block revision on the case.
- **S3b (Choice)**: RECOGNISE (to **S4a**), joint REVISE (to **S4b**), or WAIT (self-loop).
- **S4a (Recognised)**: remedy active + $\mathsf{Imm}(R)$. In **S1**, **S4a** is discharged via settlement.
- **S4b (Revised)**: substitute bundle created; conflicts deactivated under mutex; $\mathsf{Imm}(V)$ on the old case; loop back to **S0** so multiple, purpose-preserving revisions are possible.
- **S5 (Exception)**: $\mathsf{Imm}(V)$; recognition disabled; consensual revision still possible to preserve purpose, or the contract can be terminated.

Table 3 names the transitions and separates guards from acts. Fact-transitions are *performance* (S0→S1), *issue* (S0→S2), *exception* (S2→S5). Act-transitions are RECOGNISE (S3b→S4a), joint REVISE (S3b→S4b; S5→S4b), WAIT (S3b→S3b), *settlement* (S3a/S4a→S1), and *termination* (S5→S1). A configuration flag sets remedies as *alternative* (the primary is discharged once a remedy is activated) or *cumulative* (the primary remains until the issue is resolved); mutex guards prevent contradictory duties. Exceptions differ from licences: a true exception creates $\mathsf{Imm}(V)$ and excludes sanctions; a mere licence may suspend a primary but, when it lapses, recognition is not disabled by that licence alone. RoD transitions depend on authorisation (who may exercise the re-

| ID | Name | What it means | Bundle highlights |
|---|---|---|---|
| S0 | Primaries active | Preconditions met; no mutex conflict. | Claims induce duties; revision power available (if configured). |
| S1 | Fulfilled / discharge | A one-shot element is performed. | Discharge that particular claim/duty; next period/case. |
| S2 | Case open (non-fulfilment) | Due date passes or performance fails. | CTD-Claim: auto remedy; CTD-Power: choice emerges. |
| S3a | Remedy active (auto) | Auto CTD-claim detached. | Remedy live + $\mathsf{Imm}(R)$ (blocks revision). |
| S3b | Choice (CTD-Power) | No exception; CTD-Power regime. | RECOGNISE / REVISE (joint) / WAIT. |
| S4a | Recognised | Recognition exercised. | Remedy live + $\mathsf{Imm}(R)$. |
| S4b | Revised | Joint revision created. | New primaries; $\mathsf{Imm}(V)$; loop to S0. |
| S5 | Exception | True exception. | $\mathsf{Imm}(V)$; no sanctions. |

**Table 2.** Rights system states. Each state is a *bundle condition* indicating which claims, powers, and immunities exist now and why.

| From → To | Trigger (guard ∧ act) | Effect on bundle | Notes |
|---|---|---|---|
| S0 → S1 | Performance | Discharge performed claim/duty. | Next period/case → S0. |
| S0 → S2 | Issue | Open a case. | Regime/exception evaluated. |
| S2 → S3a | CTD-Claim ∧¬Exc | Activate remedy + $\mathsf{Imm}(R)$. | No agent choice. |
| S2 → S3b | CTD-Power ∧¬Exc | Offer recognise / revise / wait. | Choice point. |
| S2 → S5 | Exception | Create $\mathsf{Imm}(V)$. | Recognition disabled. |
| S3a/S4a → S1 | Remedy settled | Discharge remedy. | Case closes. |
| S3b → S3b | WAIT | Bundle unchanged. | Self-loop. |
| S3b → S4a | RECOGNISE | Remedy + $\mathsf{Imm}(R)$. | Blocks revision on this case. |
| S3b/S5 → S4b | REVISE (as authorised) | Substitute bundle + $\mathsf{Imm}(V)$. | Loop to S0. |
| S4b → S1 | Aubstitute fulfilled | Discharge substitute. | Then S0. |
| S5 → S1 | Purpose failed | Terminate the contract. | Then S0. |

**Table 3.** Named transitions. Facts move the case; acts exercise powers. Immunities enforce per-case CTD/RoD exclusivity and document why a path is closed.

vision power) rather than on exception per se. Exceptions explain why sanction is not appropriate and so justify revision, but the power to revise is conferred by the agreement.

Figure 3 is read in three threads aligned with Section 2. *CTD-Claim* is automatic: from **S2**, the remedy becomes active (**S3a**) and revision is blocked by $\mathsf{Imm}(R)$. Settlement leads to **S1**, then the "next period/case" edge returns to **S0**, representing multiple CTD episodes *across* cycles. *CTD-Power* is discretionary: **S3b** presents a real choice — RECOGNISE to **S4a**, joint REVISE to **S4b**, or WAIT. *RoD* is joint and forward-looking: **S4b** loops to **S0** with a new bundle, which enables targeted amendments over time while provenance and mutex constraints remain auditable. Case identity is clear: each due item (or declared issue) opens exactly one case; exclusivity is case-specific, and each case cycles naturally across periods. Concurrency is handled institutionally: the first committed act is protected and subsequent conflicting acts are rendered inadmissible.

In sum, the rights system operationalises Sections 1–2. CTD-*Claim* involves no choice. CTD-*Power* creates the discretionary space where RECOGNISE, REVISE, and WAIT genuinely compete on the same case. *RoD* resets the bundle while preserving the old track. The result is a short, auditable path from facts and recorded acts to live rights and duties, with invariants and exclusivity guaranteed by design.

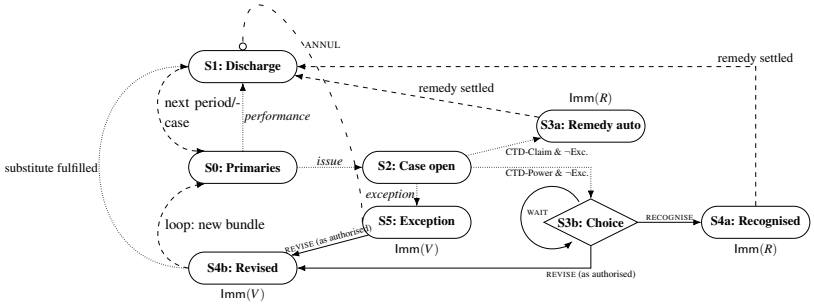



**Figure 3.** The rights system in a compact state machine. CTD-*Claim* offers no choice; CTD-*Power* offers a choice. *RoD* loops to **S0** (with a new bundle). $\mathsf{Imm}(R)$ blocks revision on the case; $\mathsf{Imm}(V)$ blocks recognition. Dotted lines denote facts, dashed lines the configuration, and solid lines legal acts.

## 4. Collaborative Agentic AI System for Rights-Based Agreements

Section 3 showed that a rights-first model suffices for the *institutional* system: the rights system computes, from facts and recorded acts, which claims and powers are admissible at each stage. We now show that the same approach suffices for an *agent* system concerned with practical reasoning. Taken together, the institutional and agent systems form a multi-agent system: the rights system provides admissible options while the agent system selects one of them for execution, both governed by the logical constraints in the specification. This is our second proof of concept: an *agentic AI in Law* can operate within the rights-first framework by adding a transparent, reason-based choice mechanism for exercising discretionary powers and producing a single executable act.

We use *agentic AI* in the general sense of systems able to set goals, plan, call external tools, coordinate with others, and maintain working memory, often implemented with Large Language Model (LLM) components [16]. In our setting, an agentic AI is an *autonomous program* representing a contractual party. It interfaces with the rights system and exercises discretion by selecting from admissible acts. LLM components may assist as personal assistants, e.g. retrieving contractual grounds, organising evidence, or drafting auditable explanations, but the assignment of normative weights remains a matter of policy, precedent, and evidence.

To explain how an agent selects from admissible acts, we adopt Tucker's *dual-scale theory of reasons* [17]. A single-scale picture collapses all reasons into one number and presumes a fixed exchange between "counts-in-favor" and "makes-obligatory", which distorts normative phenomena. Dual scale separates these forces. Each agent identifies relevant grounds (liquidity, fairness, reputation) and activates them contrastively: a ground becomes a reason for $a$ only in relation to an alternative $b$, and its weight may vary across comparisons (*contrastivism* [17, p. 122]). Reasons come with three weight values: (i) *justifying weight (JW)*, the degree to which an act is pushed towards permissibility; (ii) *requiring weight (RW)*, the degree to which an act is required, pushing alternatives towards impermissibility; and (iii) *commending weight (CW)*, the degree to which one permissible act is morally preferable to another: "altruistic benefits always commend more than self-interested benefits of the same size" [17, p. 172]. The Permission Scale checks whether one JW outweighs the RW of its rivals, while the Commitment Scale mirrors it by flipping sides. A requirement arises when an act is permissible and all alternatives are impermissible.

*Formal rules.* For an agreement $\iota$, $\mathscr{A}_\iota = \{\text{RECOGNISE}, \text{WAIT}\} \cup \{\text{REVISE}(r) \mid r \in R_\iota\}$ denotes the set of admissible acts. For any ordered pair $(a,b)$ of distinct acts, let $\mathscr{G}_\iota(a \parallel b)$ be the contrast-indexed set of grounds that support act $a$ against act $b$. Each $g \in \mathscr{G}_\iota(a \parallel b)$ contributes one or more reasons with weight $w_g \in \mathbb{N}$ ranging over natural numbers:

$$\mathsf{JW}(a \parallel b) = \sum_{g \in \mathscr{G}_\iota^{\text{just}}(a \parallel b)} w_g, \quad \mathsf{RW}(a \parallel b) = \sum_{g \in \mathscr{G}_\iota^{\text{req}}(a \parallel b)} w_g, \quad \mathsf{CW}(a) = \sum_{g \in \mathscr{G}_\iota^{\text{com}}(a)} w_g.$$

Dual-scale verdicts are pairwise: $a$ is *permissible* against $b$ iff $\mathsf{JW}(a \parallel b) \geq \mathsf{RW}(b \parallel a)$, has a *commitment* against $b$ iff $\mathsf{RW}(a \parallel b) > \mathsf{JW}(b \parallel a)$, and is *required* if both hold. With multiple options, permissibility and requirement mean winning all pairwise contests. Commending weight is not used to determine the deontic status of acts but only for ranking within the permissible set, thus refining choice (akin to advice or soft-law guidance). If no act is permissible, the default is WAIT.

*Running example (Public Park Contract).* Contractor $A$ delivered, $M$ missed timely payment. Admissible acts: RECOGNISE (REC), WAIT, REVISE(30d) (REV30D), and REVISE(14d) (REV14D). We show three contrasts. Each row lists one active and contrast-indexed reason for the *left* option with its $(jw, rw)$. In Contrast 3, where we have two tied permissible acts, commending weight comes in to make the decision.

**Contrast 1:** REVISE(30d) vs. RECOGNISE

| $(g, a \,\|\| b)$ | Justifying weight | Requiring weight | Comment |
|---|---|---|---|
| (purpose, REV30D $\|\|$ REC) | 2 | 1 | preserve purpose; avoid closure |
| (liquidity$_A$, REV30D $\|\|$ REC) | 2 | 1 | utility/efficiency for $A$ |
| (reputation$_M$, REV30D $\|\|$ REC) | 1 | 2 | sanction harms $M$, reputation/compliance |
| (compliance, REC $\|\|$ REV30D) | 1 | 3 | policy duty |
| (liquidity$_A$, REC $\|\|$ REV30D) | 2 | 0 | immediate liquidity |

*Totals:* JW(REV30D $\|$ REC) $= 5 \geq$ RW(REC $\|$ REV30D) $= 3$, so REV30D is permissible; RW(REV30D $\|$ REC) $= 4 \geq$ JW(REC $\|$ REV30D) $= 3$, so REV30D has a commitment. Thus, rev30d is required. Meanwhile, JW(REC $\|$ REV30D) $= 3 <$ RW(REV30D $\|$ REC) $= 4$, so REC is not permissible; and RW(REC $\|$ REV30D) $= 3 <$ JW(REV30D $\|$ REC) $= 5$, so REC is not a commitment. Thus, REC is not required.

**Contrast 2:** REVISE(30d) vs. WAIT

| $(g, a \,\|\| b)$ | Justifying weight | Requiring weight | Comment |
|---|---|---|---|
| (opCont, REV30D $\|\|$ WAIT) | 2 | 1 | keep services, utility |
| (budget, REV30D $\|\|$ WAIT) | 1 | 1 | administrative timing |
| (uncertainty, WAIT $\|\|$ REV30D) | 1 | 0 | information-gathering, uncertainty |

*Totals:* JW(REV30D $\|$ WAIT) $= 3 \geq$ RW(WAIT $\|$ REV30D) $= 0$, so REV30D is permissible; RW(REV30D $\|$ WAIT) $= 2 \geq$ JW(WAIT $\|$ REV30D) $= 1$, so REV30D has a commitment. Thus, REV30D is required. Meanwhile, JW(WAIT $\|$ REV30D) $= 1 <$ RW(REV30D $\|$ WAIT) $= 2$, so WAIT is not permissible; and RW(WAIT $\|$ REV30D) $= 0 <$ JW(REV30D $\|$ WAIT) $= 3$, so no commitment holds for WAIT. Thus, WAIT is not required.

**Contrast 3:** REVISE(14d) vs. REVISE(30d)

| $(g, a \,\|\| b)$ | Justifying weight | Requiring weight | Comment |
|---|---|---|---|
| (fairness$_A$, REV14D $\|\|$ REV30D) | 3 | 1 | fairness to $A$, cooperation |
| (liquidity$_A$, REV14D $\|\|$ REV30D) | 2 | 0 | faster cash, utility |
| (cashflow$_M$, REV30D $\|\|$ REV14D) | 1 | 0 | $M$'s short-term stress, utility |

*Totals:* JW(REV14D $\|$ REV30D) $= 5 \geq$ RW(REV30D $\|$ REV14D) $= 0$, so REV14D is permissible; RW(REV14D $\|$ REV30D) $= 1 =$ JW(REV30D $\|$ REV14D) $= 1$, so REV14D is not required. Meanwhile, JW(REV30D $\|$ REV14D) $= 1 \geq$ RW(REV14D $\|$ REV30D) $= 1$, so REV30D is permissible. But RW(REV30D $\|$ REV14D) $= 0 <$ JW(REV14D $\|$ REV30D) $= 5$, so no commitment holds for REV30D. Both are permissible. However, commending weight favours fairness, which counts as an altruistic benefit that tips the choiceworthiness balance.

Given this structure, the agent system can be viewed as a collaborative decision pipeline (Fig. 4) that operationalises the dual-scale model for individual reasoning and shows how multiple agent AIs interact to select a single admissible act.

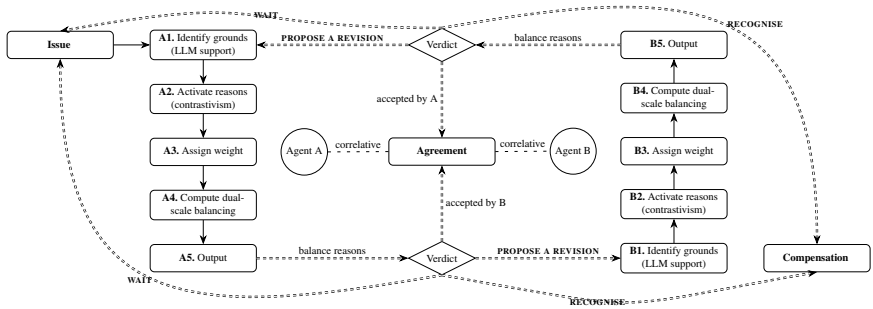



**Figure 4.** Collaborative agentic AI pipeline for rights-based agreements. Each party is represented by an agent that (1) identifies grounds (with LLM support), (2) activates reasons via contrastivism, (3) assigns weights from policy, precedent, and evidence, (4) computes dual-scale totals, and (5) outputs an act — required, best commended permissible, or WAIT if none are permissible. For verdicts, agents negotiate proposals with mutual acceptance, while recognition and waiting remain unilateral. The diagram shows their revision–accept loop yielding legally intelligible outcomes.

## 5. Related work

Contrary-to-Duty (CTD) reasoning is central to modelling norm compliance, especially in legal informatics [4, 5]. The traditional *duty-first* view motivates rich logical toolkits — dyadic deontic logic [3, 18], temporal/dynamic deontic logics [19, 20], and non-monotonic approaches (e.g. argumentation [21, 14], constrained I/O logic [22, 23]). Yet, this paradigm only partially reflects the depth of Hohfeld's legal framework [6, 13, 24].

We instead adopt a *rights-first* stance, where claims and powers stand as irreducible primitives. This opens an *operational* perspective on CTD reasoning, akin to system design, which clarifies the *meta*-conditions of compliance — when and which rights are exercised — a dimension missing from duty-first logics.

Our perspective bridges AI and Law by shifting from "*logics about norms*" to "*systems with norms*". Prior work on organisations [25, 19] and agreement technologies [26, 27, 7] model inter-agent complexity via normative multi-agent systems (MAS) [28, 8] and argumentation-based negotiation [29, 30]. Unlike accounts that separate agent-dependence [31] from compliance [32], our specification integrates them via Hohfeld's correlative rights, shifting the focus from *agents* to *rights* and linking the *institutional* (compliance) and *agentic* (trust, reputation, choice) layers at the systems level.

Recent work on *agentic AI* in law addresses tasks from extracting normative statements [33] to supporting knowledge representation [34] and dispute resolution [35] using LLMs. In this light, normative MAS and agreement technologies are not only a conceptual bridge between logics and institutional systems but also a foundation for designing legally intelligible AI agents. Our rights-first approach complements this by assigning LLMs a bounded role — assisting recall, drafting, and explanation — while supporting formal theory and safe integration into legal and contractual systems.

## 6. Summary

This paper introduces a *rights-first* theoretical framework for agreement lifecycle management that *reconceptualises CTD reasoning*. We take a design-plus-analysis approach: alongside formal study, we build deontic systems that are both executable and auditable. Our focus is on the most persistent challenge — contrary-to-duty (CTD) reasoning. By re-reading CTD reasoning through Hohfeld's lens, we separate (i) *CTD-Claim* (automatic remedial detachment), (ii) *CTD-Power* (recognition-based remedial detachment), and (iii) *Revision-of-Duty* (RoD, purpose-preserving revision), which enables us to design an operational pipeline to implement CTD-based reasoning as an agentic AI system.

Future work includes implementing *rights-based Answer Set Programming (ASP) automation* [15, 36] to align CTD rights with Hohfeldian bundles in practical reasoning [14, 37], and analysing semantics using Deontic Equilibrium Logic [36] to model the modification of rights under formal constraints. We will also study how knowledge and beliefs shape CTD rights, determining when non-fulfilments count as violations, exceptions, or cooperative revisions. A fault-tolerant rights system [28] will ensure robustness under uncertainty while preserving consistency, fairness, and provenance. Another natural step is to move beyond the proof of concept and develop a fuller account of how agentic AI in Law can operate within the rights-first framework: with a specification of how LLMs assist in identifying grounds and reasons, and how agents negotiate using symbolic reasoning and LLMs [38] for revision.

**Acknowledgements**

We are grateful to Château de Rastignac, where our work first began. We also thank the Luxembourg National Research Fund (FNR) for its support through the project Deontic Logic for Epistemic Rights (DELIGHT) (OPEN O20/14776480). Huimin Dong is supported by the Austrian Science Fund (FWF) and the FNR through the LoDEX project [doi: 10.55776/I6372 and INTER/DFG/23/17415164/LODEX]. Leendert van der Torre acknowledges the financial support of the FNR through the following projects: The Epistemology of AI Systems (EAI) (C22/SC/17111440), DJ4ME – A DJ for Machine Ethics: the Dialogue Jiminy (O24/18989918/DJ4ME), Logical Methods for Deontic Explanations (LoDEx) (INTER/DFG/23/17415164/LoDEx) and Symbolic and Explainable Regulatory AI for Finance Innovation (SERAFIN) (C24/19003061/SERAFIN). Liuwen Yu thanks the FNR for their support through the SERAFIN project (C24/19003061/SERAFIN) and the University of Luxembourg for the Marie Speyer Excellence Grant for the project Formal Analysis of Discretionary Reasoning (MSE-DISCREASON).

**References**

[1] Chisholm RM. Contrary-to-Duty Imperatives and Deontic Logic. Analysis. 1963;24(2):33-6.

[2] Prakken H, Sergot M. Contrary-to-duty obligations. Studia Logica. 1996;57(1):91-115.

[3] van der Torre L, Tan YH. Contrary-to-duty reasoning with preference-based dyadic obligations. Annals of Mathematics and Artificial Intelligence. 1999;27(1):49-78.

[4] Hage J. Contrary to Duty Obligations A Study in Legal Ontology. In: Proceedings of JURIX 2001; 2001. p. 89-102.

[5] Bench-Capon TJ. Transition systems for designing and reasoning about norms. Artificial Intelligence and Law. 2015;23(4):345-66.

[6] Hohfeld WN. Some Fundamental Legal Conceptions as Applied in Judicial Reasoning. Yale Law Journal. 1913;23(1):16-59.

[7] Ossowski S, editor. Agreement Technologies. Berlin, Heidelberg: Springer; 2013.

[8] Chopra A, van der Torre L, Verhagen H, Villata S, editors. Handbook of Normative Multiagent Systems. College Publications; 2018.

[9] Carmo J, Jones AJ. Deontic logic and contrary-to-duties. In: Handbook of Philosophical Logic: Volume 8. Springer; 2002. p. 265-343.

[10] Hart HL. Discretion. Harvard Law Review. 2013;127(2):652-65.

[11] Gabbay D, Horty J, Parent X, van der Meyden R, van der Torre L, editors. Handbook of Deontic Logic and Normative Systems. College Publications; 2013 (v. 1), 2021 (v. 2).

[12] Harel D, Kozen D, Tiuryn J. Dynamic logic. MIT Press; 2000.

[13] Markovich R. Understanding Hohfeld and formalizing legal rights: the Hohfeldian conceptions and their conditional consequences. Studia Logica. 2020;108(1):129-58.

[14] Atkinson K, Bench-Capon T. Practical reasoning as presumptive argumentation using action based alternating transition systems. Artificial Intelligence. 2007;171(10-15):855-74.

[15] Azzopardi S, Pace GJ, Schapachnik F, Schneider G. Contract automata: an operational view of contracts between interactive parties. Artificial Intelligence and Law. 2016;24(3):203-43.

[16] Acharya DB, Kuppan K, Divya B. Agentic AI: Autonomous intelligence for complex goals–a comprehensive survey. IEEE Access. 2025.

[17] Tucker C. The Weight of Reasons: A Framework for Ethics. New York: Oxford University Press; 2025.

[18] Ciabattoni A, Olivetti N, Parent X. Dyadic obligations: proofs and countermodels via hypersequents. In: Proceedings of PRIMA 2022. Springer; 2022. p. 54-71.

[19] Governatori G, Rotolo A, Sartor G. Temporalised normative positions in defeasible logic. In: Proceedings of ICAIL 2005; 2005. p. 25-34.

[20] Meyer JJC, et al. A different approach to deontic logic: deontic logic viewed as a variant of dynamic logic. Notre Dame J Formal Log. 1988;29(1):109-36.

[21] Bench-Capon TJM. Persuasion in Practical Argument Using Value-Based Argumentation Frameworks. Journal of Logic and Computation. 2003;13(3):429-48.

[22] Makinson D, Van Der Torre L. Constraints for input/output logics. Journal of philosophical logic. 2001;30(2):155-85.

[23] van Berkel K, Markovich R, Straßer C, van der Torre L. Arguing About Choosing a Normative System: Conflict of Laws. In: Sileno G, Spanakis J, van Dijck G, editors. Proceedings of JURIX 2023. vol. 379 of Frontiers in Artificial Intelligence and Applications. IOS Press; 2023. p. 73-82.

[24] Dong H, Roy O. Dynamic logic of legal competences. Journal of Logic, Language and Information. 2021;30(4):701-24.

[25] Jones AJI, Sergot M. A Formal Characterisation of Institutionalised Power. Journal of the IGPL. 1996;4(3):429-45.

[26] Boella G, Pigozzi G, van der Torre L. Five Guidelines for Normative Multiagent Systems. In: Governatori G, editor. Proceedings of JURIX 2009. vol. 205 of Frontiers in Artificial Intelligence and Applications. IOS Press; 2009. p. 21-30.

[27] Boella G, van der Torre L. Reasoning for Agreement Technologies. In: Raedt LD, Bessiere C, Dubois D, Doherty P, Frasconi P, Heintz F, et al., editors. Proceedings of ECAI 2012. vol. 242 of Frontiers in Artificial Intelligence and Applications. IOS Press; 2012. p. 895-6.

[28] Boella G, van der Torre L. Constitutive Norms in the Design of Normative Multiagent Systems. In: Proceedings of CLIMA 2005. Springer; 2005. p. 303-19.

[29] Amgoud L, Kaci S. On the generation of bipolar goals in argumentation-based negotiation. In: Proceedings of ArgMAS 2004. Springer; 2004. p. 192-207.

[30] Liao B, Pardo P, Slavkovik M, van der Torre L. The Jiminy Advisor: Moral Agreements among Stakeholders Based on Norms and Argumentation. Journal of Artificial Intelligence Research. 2023;77:737-92.

[31] Boella G, Caire P, van der Torre L, Villata S. Dependence Networks for Agreement Technologies. In: Ossowski S, Toni F, Vouros GA, editors. Proceedings of AT 2012. vol. 918 of CEUR Workshop Proceedings. CEUR-WS.org; 2012. p. 109-10.

[32] Marosin D, Proper HA, van der Torre L. Changing Agreements: Intention Reconsideration based on Assumptions and Reasons. In: Ossowski S, Toni F, Vouros GA, editors. Proceedings of AT 2012. vol. 918 of CEUR Workshop Proceedings. CEUR-WS.org; 2012. p. 262-3.

[33] Zin MM, Satoh K, Borges G. Leveraging LLM for Identification and Extraction of Normative Statements. In: Proceedings of JURIX 2024. IOS Press; 2024. p. 215-25.

[34] Billi M, Pisano G, Sanchi M. Fighting the knowledge representation bottleneck with Large Language Models. In: Proceedings of JURIX 2024. IOS Press; 2024. p. 14-24.

[35] Pawar SS, Apte M, Palshikar GK, Ali B, Ramrakhiyani N. DRAssist: Dispute Resolution Assistance using Large Language Models. In: Proceedings ICAIL 2025. forthcoming. ACM; 2025. .

[36] Cabalar P, Ciabattoni A, van der Torre L. Deontic equilibrium logic with explicit negation. In: Proceedings of JELIA 2023; 2023. p. 498-514. Springer.

[37] Dong H, Doder D, Li X, Markovich R, van der Torre L, van Zee M. Rights and Practical Reasoning in Deontic Logic. In: Proceedings of DEON 2023. College Publications; 2023. p. 1-19.

[38] Yu L, Liga D, Markovich R. Addressing the Right to Explanation and the Right to Challenge through Hybrid-AI: Symbolic Constraints over Large Language Models via Prompt Engineering. In: Proceedings ICAIL 2025. forthcoming. ACM; 2025.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251578*

# Does ChatGPT Understand the Law? A Case Study on Road Homicide in Italy

Grazia GARZO [a,1], Alessandro PALUMBO [b]
[a] *University of Siena, Italy*
[b] *CentraleSupélec, Inria, CNRS, IRISA, France*

**Abstract.** This manuscript proposes a structured methodology and a replicable framework for empirically assessing the argumentative and interpretative capabilities of Large Language Models (LLMs) in the field of justice. As a demonstrative case, the framework is applied to the specific offence of vehicular homicide under Italian law, using GPT-4o as the tested model.

The evaluation is structured in two complementary phases. The first phase investigates the model's conceptual understanding in isolation: 60 legal concepts were tested through targeted prompts eliciting definitions, legal nuances, and illustrative examples. The responses were scored to assess the model's abstract comprehension of core legal notions. The second phase evaluates the model's ability to recognize, interpret, and apply these same legal concepts within real judicial reasoning. The model was provided with complete rulings from the Italian Court of Cassation and asked to summarize the decisions, identify the ratio decidendi, and reconstruct the legal reasoning underlying the Court's conclusions.

Outputs from both phases were assessed by a legal expert, who evaluated coherence, conceptual depth, and interpretative accuracy. The results highlight some correlation between the model's prior conceptual grounding and its ability to understand and replicate complex judicial reasoning. These findings underscore the importance of expert oversight in any forensic or judicial application of LLMs.

Beyond the specific case of vehicular homicide, the study proposes a generalizable framework for evaluating the legal reasoning capabilities of AI systems across different domains.

**Keywords.** Generative Artificial Intelligence, Large Language Models, Legal Understanding, Criminal Law, Road Homicide

## 1. Introduction and Motivation

In recent years, generative artificial intelligence has assumed a prominent role in debates on the future of legal scholarship, emerging as a transformative technology for the analysis, interpretation, and even application of the law [1,2]. Large Language Models (LLMs) such as GPT have distinguished themselves by their ability to generate coherent texts, answer complex questions, and simulate argumentative reasoning in natural language [3,4]. Their apparent competence, even in specialized domains like medicine, law, and public administration [5,6], raises a pressing question: do these systems genuinely understand legal concepts, or do they simply reproduce statistical correlations?

[1] Corresponding Author: Grazia Garzo, g.garzo@student.unisi.it

This question is especially relevant in practical contexts where LLMs are deployed: drafting legal documents, providing automated advice, simplifying regulations, or analyzing case law [7]. Some jurisdictions have already introduced AI tools to support predictive jurisprudence [8], while others are evaluating their use in procedural and administrative domains [9,10]. In such scenarios, conceptual accuracy becomes crucial: if a model fails to distinguish between negligence and intent, or between causation and intervening causes, its answers may compromise fundamental principles of legality.

The Italian offence of vehicular homicide (Art. 589-bis c.p.), introduced by Law No. 41/2016, provides an emblematic case study. Unlike involuntary manslaughter (Art. 589 c.p.), it introduces harsher penalties, a detailed typification of conduct, and an aggravated structure aimed at punishing particularly dangerous driving [11,12]. The law presumes heightened culpability when the driver is impaired, narrowing judicial discretion in assessing negligence. Scholars have noted tensions with the principles of proportionality and personal responsibility [12].

Causality further complicates judicial evaluation: courts must apply the "more likely than not" standard, the lawful alternative conduct test, and analyze possible intervening causes. In multi-actor cases, liability depends on a differentiated assessment of contributory negligence and foreseeability of the outcome [11]. Another delicate issue is the distinction between conscious negligence and conditional intent, particularly in cases of egregious misconduct (street racing, driving with a revoked license, repeated violations). The Court of Cassation has repeatedly stressed that mere awareness of risk is not intent unless the agent accepts the possibility of the harmful outcome.

These features make vehicular homicide an ideal testbed for assessing whether LLMs can:

- Distinguish subjective and objective legal configurations;
- Correctly apply causal attribution standards;
- Differentiate aggravated negligence, conscious negligence, and conditional intent;
- Understand the implications of aggravating factors.

The degree of consistency in model outputs within this dense normative framework allows us to evaluate not only linguistic competence but also the capacity to generate logically grounded legal reasoning. This study explores three key dimensions: the influence of conceptual familiarity on legal reasoning; the recurrence of errors in applying criminal law concepts; and the possibility of constructing a replicable empirical framework for evaluating computational legal understanding. Unlike much of the existing literature, which focuses on predictive accuracy, this work emphasizes conceptual depth, jurisprudential grounding, and qualitative error taxonomies [6,13,14]. The contribution is twofold: it proposes a replicable methodology for assessing legal reasoning in LLMs, and it provides evidence of structural limitations that underscore the indispensable role of human experts in safeguarding legality and proportionality.

The remaining part of the manuscript is organised as follows: Section 2 reviews the main contributions in the literature concerning the legal capabilities of LLMs, with particular attention to the distinction between predictive applications and reasoning-oriented evaluations. Section 3 introduces the experimental methodology, outlining the structure and rationale of the two-phase framework and describing the procedures used to elicit and evaluate the model's legal reasoning. Section 4 presents the empirical results obtained from the application of this framework to GPT-4o, highlighting key patterns in

the model's performance, correlations between conceptual competence and argumentative behavior, and recurring errors. Section 5 discusses the broader implications of these findings, including methodological, ethical, and legal considerations. Finally, Section 6 concludes the paper by summarizing the main contributions.

## 2. Related Work

In recent years, the use of LLMs has attracted growing interest within both the legal and computer science research communities, generating a vast body of literature exploring the potential and limitations of artificial intelligence in the legal domain [15]. Two main research directions have emerged: on the one hand, the study of the predictive capabilities of the models (*legal prediction*); on the other, the analysis of their conceptual and argumentative competence (*legal reasoning*).

From a predictive standpoint, Zhong et al. developed an automated judgment prediction system within the Chinese legal context, based on deep neural networks and semantic representations of facts. The system achieved promising accuracy but struggled to generalize to borderline cases [16]. Chalkidis et al. applied LSTM architectures and transformer models to predict legal labels in cases before the European Court of Human Rights, reporting difficulties with complex or atypical cases [17]. Medvedeva et al. conducted a systematic comparison of various machine learning models applied to Dutch case law, noting a decline in performance on tasks requiring causal reasoning [18]. Aletras et al. proposed an SVM-based predictive model for ECtHR cases, demonstrating that models tend to exploit lexical correlations rather than grasp the underlying legal rules [19]. The study in [20] presents a predictive approach to identifying vehicular homicide using AI techniques to extract relevant features from crime scenes. Based on Boolean logic, the system anticipates legal outcomes with an accuracy of 83.2%. A further development is illustrated in [21], which introduces the first hardware circuit capable of replicating judicial reasoning, achieving an accuracy of 86%.

In terms of legal reasoning, Bommarito and Katz tested GPT-3.5 on the U.S. bar exam, finding high performance in civil law questions but significant instability in substantive criminal law [3]. Zheng et al. introduced CaseHOLD, a multiple-choice benchmark based on judicial decisions, highlighting that models tend to rely on superficial linguistic patterns rather than normative comprehension [22]. Nanda et al. emphasized the issue of *fairness* in legal question answering systems, showing a disconnection between surface-level accuracy and argumentative adequacy [23]. Dahl et al. documented recurring *hallucination* phenomena in LLMs, including the generation of fictitious legal or jurisprudential references [6].

To support the empirical evaluation of the legal capabilities of language models, numerous benchmarks have been proposed. Hendrycks et al. introduced MMLU, a multidisciplinary test that includes legal questions [4]. Zheng et al. published LegalBench, a suite of legal reasoning benchmarks [24]. Chalkidis et al. presented LexGLUE, a collection of NLP legal datasets for various tasks [25]. Shu et al. evaluated GPT-4 on legal reasoning tasks within the LawLLM project, highlighting both its strengths and its limitations in handling complex cases [5]. Li et al. developed a knowledge-based approach, K-LJP, to predict relevant charges and articles in Chinese legal judgment prediction [26]. Rabelo et al. curated the Japanese COLIEE benchmark for legal entailment and retrieval tasks [27].

Despite the breadth of these contributions, many benchmarks adopt multiple-choice formats that neglect the evaluation of argumentative structure and favor answers based on lexical similarity. Guha et al. showed how simplified prompts can lead to an overestimation of models' argumentative capacities [10,10]. Blair-Stanek et al. tested GPT-3.5 and GPT-4 on complex tasks, revealing difficulties in maintaining causal coherence and normative consistency in their explanations [28].

The literature has also addressed the ethical and regulatory implications of using AI in the legal domain. Livermore and Rockmore analyzed the risk of propagating jurisprudential bias, emphasizing the need for human oversight in generative systems [7]. Hurshman warned against the opacity of training data and the challenges of model auditability [29]. Zhang et al. proposed an innovative ethical evaluation framework specifically designed to test LLM capabilities in legal contexts, considering not only knowledge but also ethical requirements [30]. Isaac et al. advocated for a legal regulation of AI systems grounded in transparency, traceability, and legal accountability [31]. Garzo et al. explored the legal and ethical implications of integrating predictive AI technologies into the Italian criminal justice system, with a particular focus on decision transparency, judicial autonomy, and responsibility [32]. Recent approaches have also explored the use of controlled languages for the formalization of legal norms. In [33] the authors propose a human-in-the-loop methodology for translating EU legal provisions and case law into Attempto Controlled English, contributing to the debate on legal formalization and computable law.

In the European context, Luciano Floridi has outlined a model of *Trustworthy AI* aimed at the responsible integration of AI into legal systems, with emphasis on human oversight and explainability of decisions [34]. Giovanni Sartor analyzed the tension between algorithmic efficiency and the principle of legality, raising concerns about the compatibility of automated decision-making with procedural guarantees in civil law systems [9]. In the state of the art, there are areas that still leave room for methodological improvement, particularly with regard to how legal LLMs' reasoning is evaluated. This study addresses three specific dimensions:

1. It proposes a general-purpose conceptual framework to assess whether (and in what way) LLMs understand and apply fundamental principles of criminal law;
2. It complements existing evaluation benchmarks (typically quantitative and rooted in English-language common law traditions) with an approach tailored to the normative structures of civil law systems;
3. It offers a structured classification of conceptual errors, enabling a more precise analysis of the epistemic limitations and reasoning behaviors exhibited by the model.

## 3. Methodology

In order to rigorously and replicably assess the ability of the ChatGPT-4o model to understand, manipulate, and apply complex legal concepts within the framework of Italian criminal law–with particular reference to the offence of vehicular homicide under Article 589-bis of the Penal Code–this study adopts a two-phase experimental protocol. The methodological goal is to move beyond traditional benchmarks based on predictive accuracy or surface-level performance, and instead to propose an empirical framework

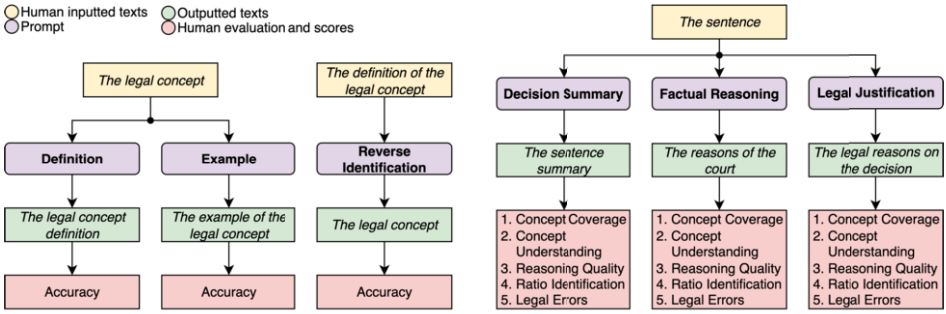



**Figure 1.** The workflows: Phase 1 (left); Phase 2 (right).

capable of evaluating the depth of conceptual understanding, the consistency of legal reasoning, and the referential soundness of the model's outputs.

The proposed approach is grounded in the distinction between two complementary dimensions of legal understanding, both examined in relation to concepts central to the offence of vehicular homicide under Italian criminal law. These dimensions are operationalized through a two-phase experimental methodology, with each phase designed to evaluate one aspect of the LLM's legal competence:

1. **Stand-alone Conceptual Competence:** testing the model's ability to define, explain, and recognize criminal law concepts in isolation, without contextual anchoring in specific cases;
2. **Contextual Application in Jurisprudential Reasoning:** assessing the model's capacity to apply those same concepts within complex argumentative structures derived from real judicial decisions.

Figure 1 illustrates the overall structure of the two-phase experimental workflow. Each phase is articulated through a series of targeted prompts administered to the model, designed to elicit legal knowledge and reasoning. The outputs generated in response to these prompts were manually evaluated by a legal expert, using a structured scoring system. The assessment took into account the precision, clarity, and legal soundness of the responses, assigning scores that reflect the model's degree of conceptual understanding and argumentative reliability. While the present study focuses on GPT-4o and the offence of vehicular homicide, the methodology is designed to be model-agnostic and broadly applicable to other criminal offences and legal domains.

The crime of vehicular homicide was selected as the use case for applying the proposed methodology, due to its high conceptual density, legal complexity, and significance within contemporary Italian criminal jurisprudence. As an offence governed by a stratified and nuanced legal framework, it requires the interpreter to navigate intricate distinctions between negligence and intent, assess complex causal relationships, and apply aggravating factors with normative precision. These features make it an ideal benchmark for testing the model's ability to handle structured legal reasoning and to engage meaningfully with the doctrinal foundations of criminal liability.

### *3.1. Phase 1: Stand-alone Conceptual Competence*

The first phase of the experimental protocol focused on evaluating the model's ability to understand legal concepts in isolation, independently of any specific factual or jurispru-

dential context. A total of 60 key concepts were identified as central to the interpretation and application of Article 589-bis of the Italian Penal Code, which governs the offence of vehicular homicide. The selected concepts cover a broad range of substantive criminal law notions, including negligence, intent, causation, aggravating circumstances, and forms of liability, and are documented and explained in detail in Table 1.

Once the glossary was constructed, each concept was submitted to the model through three distinct prompts, designed to test different facets of conceptual understanding:

- **Definition** (`D`) **prompt:** the model was asked to define the concept in technical-legal terms;
- **Example** (`E`) **prompt:** the model was asked to provide a concrete or case-based example illustrating the concept;
- **Reverse Identification** (`RI`) **prompt:** the model was given a descriptive formulation of the concept and asked to identify which legal notion it referred to.

For each prompt, the model's response was evaluated independently by a legal expert and assigned a numerical **Accuracy** score according to the following scheme: 2 for fully correct answers (i.e., the maximum score), 1 for partially correct answers (e.g., responses containing ambiguity, simplification, or minor inaccuracies), and 0 for incorrect or misleading outputs. This three-prompt evaluation was performed for each of the 60 designated legal concepts, resulting in a total of 180 individual interactions and corresponding expert assessments.

### *3.2. Phase 2: Contextual Application in Jurisprudential Reasoning*

The second phase of the experimental protocol focuses on evaluating the model's ability to apply legal concepts within complex argumentative structures derived from real-world case law. Specifically, this phase assesses how well the model can interpret, reconstruct, and reason through actual judicial decisions concerning vehicular homicide.

To this end, a set of ten sentences from the Italian Court was previously selected based on their relevance, conceptual density, and doctrinal complexity. For each sentence, the model was presented with three distinct prompts, designed to probe different dimensions of its jurisprudential competence:

- **Decision Summary** (`DS`) **prompt:** the model was asked to summarize the content of the ruling, including the key facts and the outcome;
- **Factual Reasoning** (`FR`) **prompt:** the model was asked to explain the reasoning adopted by the Court in relation to the events that occurred and the causal chain leading to the decision;
- **Legal Justification** (`LJ`) **prompt:** the model was asked to identify and articulate the legal arguments and normative foundations that support the Court's decision.

Each response was manually evaluated by a legal expert using a structured scoring system, taking into account factors such as legal accuracy, argumentative coherence, clarity of exposition, and the correct use of legal concepts previously assessed in Phase 1. Five distinct scoring dimensions were applied to each prompt output:

1. **Concept Coverage Score:** quantifies the proportion of relevant legal concepts (out of the 60 defined in Phase 1) that are actually activated in the ruling under analysis. The score is expressed as a percentage;

2. **Concept Understanding Score:** evaluates the degree to which the model correctly interprets and applies the activated concepts in context, based on consistency with doctrine and case law. The score ranges from 0 (incorrect or incoherent use) to 2 (fully accurate and contextually appropriate);
3. **Reasoning Quality Score:** assesses the overall structure, clarity, and internal coherence of the model's response, including its argumentative soundness and ability to reconstruct judicial logic. The score ranges from 0 (poor or fragmented reasoning) to 2 (complete, clear, and logically consistent);
4. **Ratio Identification Score:** specifically measures the model's capacity to correctly identify and formulate the ratio decidendi expressed by the Court in the given ruling. The score ranges from 0 (incorrect or missing) to 2 (accurate and well-articulated);
5. **Legal Error Penalty:** a corrective factor applied in cases where the output contains significant legal inaccuracies, such as the invention of non-existent case law, misattribution of legal principles, or serious conceptual confusion. The penalty ranges from 0 (no significant error) to $-2$ (severe legal error).

The Phase 2 covered a total of 10 sentences, each subjected to three distinct prompts, resulting in 30 model outputs. As reported in Figure 1, a total of 150 individual evaluations across the full set of judicial cases were applied. The cumulative score for each sentence ranges from a theoretical minimum of $-2$ to a maximum of 7. Specifically, the maximum is obtained by: (i) achieving 100% in the *Concept Coverage Score* (i.e., 1 point); (ii) receiving a perfect score of 2 in *Concept Understanding*, *Reasoning Quality*, and *Ratio Identification*; and (iii) incurring no penalty in the *Legal Error* metric (i.e., 0). This aggregation allows for a compact yet multidimensional representation of the model's jurisprudential performance on each case analysis.

## 4. Results

All experiments were conducted in Italian, including the legal sources, conceptual mappings, and prompts. The choice of language was crucial: since the task revolved around Italian legal doctrines specifically (the road homicide crime) as well as norms from the Italian *penal code* and *highway code*, using Italian ensured conceptual and terminological accuracy. The material presented in this manuscript has been subsequently translated into English for clarity and consistency in presentation. In Phase 1, the model's knowledge of the 60 legal concepts (listed in Table 1) was evaluated along three distinct prompting tasks: definition recall (`D`), example generation (`E`), and reverse identification (`RI`). Results show that the model demonstrates a strong grasp of legal definitions, with an average score of 1.95 out of 2 in the `D` prompt. Performance drops significantly in the `E` prompt, with an average of 1, indicating difficulty in generating legally appropriate and contextually accurate examples. However, in the reverse identification task (`RI`), where the model had to recognize a concept from its description, performance improved again, reaching an average score of 1.97. Table 2 reports the results of Phase 2, in which GPT-4o was asked to analyse ten publicly available rulings from the Italian Court of Cassation concerning road homicide. Each ruling is identified by its case number and year (e.g., *74/2022*), and was evaluated across five dimensions.

Overall, the model demonstrated a not-so-small level of accuracy in identifying the legal concepts present in each ruling, with an average *coverage* score of 91.97%. The summaries generated by GPT-4o in response to the DS prompt were generally well-formed and comprehensible, with an average *Understanding* score of 2 out of 2. The model also performed well in recognising the legal reasoning embedded in these summaries, achieving an average *reasoning* score of 2. With regard to the identification of the judicial *Ratio decidendi*, the model's performance was slightly lower, with an average score of 1.8 out of 2, suggesting that it occasionally missed or only partially captured the core argumentative structure of the decisions. Penalties were applied when the output included errors or misinterpretations, leading to an average penalty score of –0.2 across all cases. For the FR prompt, the model maintained high levels of comprehension and reasoning accuracy, scoring a perfect 2 on both metrics. Additionally, it performed better than DS on the identification of the judicial ratio, with an average score of 2 and no penalties. The LJ prompt was more challenging. Although GPT-4o achieved an average score of 1.9 in understanding and maintained a strong performance in reasoning, its ability to capture the ratio of the decision dropped to 1.5. Moreover, this prompt accrued a slightly higher average penalty of –0.2.

Table 1.: List and scores of legal concepts evaluated in Phase 1

| **Concept** | **Legal Source** | D | E | RI |
|---|---|---|---|---|
| Road Homicide | Art. 589-bis c.p. | 2 | 1 | 2 |
| Serious or very serious road injuries | Art. 590-bis c.p. | 2 | 1 | 2 |
| Drunk driving | Art. 186 CdS | 2 | 1 | 2 |
| Driving under the influence of drugs | Art. 187 CdS | 2 | 1 | 2 |
| Driver fleeing the scene | Art. 189 para. 1 and 6 CdS | 2 | 1 | 2 |
| Failure to assist | Art. 189 para. 7 CdS | 2 | 1 | 2 |
| General negligence | Art. 43 c.p. | 2 | 1 | 2 |
| Specific negligence | Art. 43 c.p. | 2 | 1 | 2 |
| Conscious negligence | Case law | 1 | 1 | 2 |
| Conditional intent | Art. 43 c.p. | 2 | 1 | 2 |
| Causal link | Art. 40–41 c.p. | 2 | 1 | 2 |
| Joint participation in crime | Art. 110 c.p. | 2 | 1 | 2 |
| Aggravating factor: no license | Art. 186 para. 2-bis CdS | 2 | 1 | 2 |
| Aggravating factor: fleeing | Art. 589-bis para. 2 c.p. | 2 | 1 | 2 |
| Aggravating factor: altered state | Art. 589-bis para. 2 c.p. | 2 | 1 | 2 |
| Generic mitigating circumstances | Art. 62-bis c.p. | 2 | 1 | 2 |
| Plea bargaining | Art. 444 c.p.p. | 2 | 1 | 2 |
| Abbreviated trial | Art. 438 c.p.p. | 2 | 1 | 2 |
| Contributory negligence | Case law | 2 | 1 | 2 |
| Foreseeability of the event | Case law | 2 | 1 | 2 |
| BAC over 0.8 g/l | Art. 186 para. 2(b) CdS | 1 | 1 | 2 |
| BAC over 1.5 g/l | Art. 186 para. 2(c) CdS | 2 | 1 | 2 |
| License revocation | Art. 222 CdS | 2 | 1 | 2 |
| License suspension | Art. 218 CdS | 1 | 1 | 2 |
| Driving with revoked license | Art. 116 CdS | 2 | 0 | 2 |
| Driving without a license | Art. 116 para. 15 CdS | 2 | 1 | 2 |
| Refusal of alcohol test | Art. 186 para. 7 CdS | 2 | 1 | 2 |
| Refusal of drug test | Art. 187 para. 8 CdS | 2 | 1 | 2 |
| Mandatory arrest in flagrante delicto | Art. 380 c.p.p. | 2 | 0 | 2 |
| Detention of suspect | Art. 384 c.p.p. | 2 | 2 | 2 |
| Immediate trial | Art. 453 c.p.p. | 2 | 1 | 2 |
| Sentence upon request | Art. 444 c.p.p. | 2 | 1 | 2 |
| Conditional suspension of sentence | Art. 163 c.p. | 2 | 1 | 2 |
| Recidivism | Art. 99 c.p. | 2 | 1 | 2 |
| Dangerous driving | Art. 141 CdS | 2 | 1 | 2 |
| Excessive speed | Art. 141 para. 3 CdS | 2 | 1 | 2 |
| Failure to maintain safe distance | Art. 149 CdS | 2 | 1 | 2 |
| Running a red light | Art. 146 CdS | 2 | 1 | 2 |
| Failure to yield right of way | Art. 145 CdS | 2 | 1 | 2 |
| Aggravating factor: high BAC | Art. 589-bis para. 2 c.p. | 2 | 2 | 2 |
| Use of mobile phone while driving | Art. 173 CdS | 2 | 1 | 1 |
| Dangerous U-turn | Art. 154 CdS | 2 | 1 | 1 |
| Illegal turn | Art. 154 CdS | 2 | 1 | 2 |
| Disregard of road signs | Art. 146 CdS | 2 | 2 | 2 |
| Driving against traffic | Art. 143 CdS | 2 | 1 | 2 |
| Psychophysical impairment | Art. 187 CdS | 2 | 1 | 2 |
| Driving in dangerous conditions | Art. 187 CdS | 2 | 0 | 2 |

| Concept | Legal Source | D | E | RI |
|---|---|---|---|---|
| Alcohol test with breathalyzer | Art. 379 Reg. CdS | 2 | 1 | 2 |
| Drug test | Art. 187 CdS | 2 | 1 | 2 |
| Criminal responsibility | Art. 85 c.p. | 2 | 1 | 2 |
| Mental capacity | Art. 85 c.p. | 2 | 1 | 2 |
| State of necessity | Art. 54 c.p. | 2 | 1 | 2 |
| Self-defense | Art. 52 c.p. | 2 | 1 | 2 |
| Fortuitous event | Art. 45 c.p. | 2 | 1 | 2 |
| Mistake of fact | Art. 47 c.p. | 2 | 1 | 2 |
| Mistake of law | Art. 5 c.p. | 2 | 1 | 2 |
| Balancing of circumstances | Art. 69 c.p. | 2 | 1 | 2 |
| Accessory sanctions | Art. 19 c.p. | 2 | 1 | 2 |
| Fine | Art. 24 c.p. | 2 | 1 | 2 |
| Penalty | Art. 26 c.p. | 2 | 1 | 2 |
| **Average** | | 1.95 | 1 | 1.97 |

**Table 2.** Evaluation scores for the output of prompts in Phase 2

| Sentence | 1. Coverage | 2. Understanding | | | 3. Reasoning | | | 4. Ratio | | | 5. Penalty | | | Score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | DS | FR | LJ | DS | FR | LJ | DS | FR | LJ | DS | FR | LJ | |
| 74/2022 | 100% | 2 | 2 | 2 | 2 | 2 | 2 | 0 | 2 | 0 | −2 | 0 | 0 | 13 |
| 20253/2023 | 75% | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 0 | −1 | 14.75 |
| 5862/2021 | 100% | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 0 | 19 |
| 6366/2016 | 92.86% | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 0 | 0 | −1 | 16.93 |
| 24175/2023 | 80% | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 0 | 18.8 |
| 17209/2023 | 90% | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 0 | 18.9 |
| 50110/2019 | 100% | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 0 | 19 |
| 42099/2021 | 100% | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 0 | 19 |
| 42018/2022 | 100% | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 0 | 19 |
| 36814/2023 | 81.81% | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 0 | 18.81 |
| **Averages** | 91.97% | 2 | 2 | 1.9 | 2 | 2 | 2 | 1.8 | 2 | 1.5 | −0.2 | 0 | −0.2 | 17.72 |

## 5. Discussion

The evaluation of GPT-4o's performance in the example generation task (E) reveals a consistent pattern of partial comprehension. In almost all cases, the model was able to provide responses that were thematically relevant, but superficial, lacking the legal precision, normative anchoring, and factual coherence expected in a proper legal example. One of the most recurrent issues was the fabrication or misuse of case references: many examples included invented or misattributed citations to judicial rulings, which undermined the legal credibility of the outputs. In some cases, GPT-4o presented what appeared to be real case scenarios but attached them to incorrect or imaginary sentence numbers. In other instances, it invented illustrative examples that, while plausible, were not explicitly framed as hypothetical, thereby generating potential confusion between real jurisprudence and constructed narratives. The examples also often lacked adequate legal framing. Concepts were introduced in vague or generic terms, without specifying the legal consequences, doctrinal interpretations, or relevant codified norms. While the model showed an ability to associate the correct semantic domain with each concept, the examples it produced fell short of what would be expected from a jurist or law student trained in producing concrete, normatively grounded illustrations.

These findings highlight an important distinction between linguistic-pattern recognition and true legal competence. The model performs well in tasks that rely on identifying or reproducing familiar formulations, such as definition recall (D) and reverse identification (RI). In both cases, its strong performance appears to stem more from its ability to match lexical and syntactic patterns than from a deep conceptual understanding: GPT-4o effectively retrieves key terms and legal phrasing, producing coherent outputs without necessarily engaging in substantive legal reasoning. The situation changes in the exam-

ple generation task (E), where the model is required to construct a concrete and legally accurate illustration of the concept.

While GPT-4o demonstrated notable competence in identifying key legal concepts and reproducing elements of judicial reasoning in Italian rulings, a deeper examination of the outputs generated in Phase 2 reveals several recurring limitations. Although the model's average scores were relatively high, the accuracy and reliability of its responses were uneven when examined more closely. One recurrent issue lies in normative misclassifications. For example, in multiple instances, the model misattributed cooperative liability to Article 110 of the Penal Code, when the correct reference should have been Article 113, which governs negligent cooperation. Similarly, it occasionally introduced non-existent legal expressions (e.g., implausible causality") or employed inappropriate legal phrasing which, although colloquially intuitive, lacks technical accuracy in the Italian legal context.

Beyond structural and legal errors, the model also exhibited a number of semantic inaccuracies that significantly affect the interpretability and epistemic reliability of its outputs. Some responses were vague, incomplete, or lacked appropriate normative grounding. GPT-4o fabricated or altered judicial identifiers, associating real case numbers with fictional fact patterns, or vice versa. These forms of referential distortion not only compromise clarity but may also mislead users by attributing false legitimacy to legal arguments. Even more critically, the model at times created hallucinated legal authority, presenting entirely invented case citations as if they were real or assigning existing ones to the wrong doctrinal contexts.

## 6. Conclusion

This study presented an empirical framework for evaluating the legal reasoning abilities of Large Language Models (LLMs), with a specific focus on GPT-4o and the offence of vehicular homicide under Italian criminal law. The findings reveal that GPT-4o exhibits strong performance in definition recall and reverse identification tasks, suggesting an ability to retrieve and recognize legal formulations with high lexical fidelity. However, it struggles in tasks that require generative precision and normative contextualization, such as the construction of legal examples. Critical errors persisted, including normative misclassifications, improper references to legal sources, vague or inaccurate expressions, and occasional hallucinations. These results reinforce a crucial distinction: LLMs do not exhibit legal understanding in the human sense but rather demonstrate a form of pattern-based linguistic competence. They are far from being autonomous legal agents, but their capabilities as decision-support tools should not be underestimated. When properly supervised, they can assist in tasks such as summarization, concept retrieval, and preliminary legal analysis. Future research should explore ways to integrate these systems within legal workflows, ensuring transparency, accountability, and human oversight. From a methodological standpoint, the framework proposed in this paper offers a replicable approach for dissecting and scoring LLM outputs across conceptual and argumentative dimensions, while highlighting recurring error types that may be domain-specific. This dual-layered evaluation, grounded in legal expertise, allows for a more granular and principled assessment than traditional benchmarks based solely on predictive accuracy.

## References


[1] H. Surden. Artificial intelligence and law: An overview. *Ga. St. UL Rev.*, 35:1305, 2018.

[2] Kevin D Ashley. *Artificial intelligence and legal analytics: new tools for law practice in the digital age*. Cambridge University Press, 2017.

[3] M. Bommarito II and D. M. Katz. Gpt takes the bar exam. *arXiv:2212.14402*, 2022.

[4] D. Hendrycks et al. Measuring massive multitask language understanding. *arXiv:2009.03300*, 2020.

[5] D. Shu et al. Lawllm: Law large language model for the us legal system. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4882–4889, 2024.

[6] M. Dahl et al. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93, 2024.

[7] M. A Livermore et al. Law search in the age of the algorithm. *Mich. St. L. Rev.*, page 1183, 2020.

[8] Rachel E Stern, Benjamin L Liebman, Margaret E Roberts, and Alice Z Wang. Automating fairness? artificial intelligence in the chinese courts. *Colum. J. Transnat'l L.*, 59:515, 2020.

[9] G. Sartor. La legalità algoritmica nella giurisprudenza amministrativa. *Rivista Italiana di Informatica e Diritto*, XXV(2):1–18, 2022.

[10] N. Guha et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36:44123–44279, 2023.

[11] R. Giovagnoli. *Manuale di Diritto Penale: Parte Generale*. Itaedizioni, Roma, Italia, 2021.

[12] R. Giovagnoli. *Manuale di Diritto Penale: Parte Speciale*. Itaedizioni, Roma, Italia, 2023.

[13] J. H Choi and D. Schwarcz. Ai assistance in legal analysis: An empirical study. 2023.

[14] G. Garzo and A. Palumbo. Come l'intelligenza artificiale può supportare le decisioni giuridiche: Un caso di studio sugli omicidi stradali. *Journal of Inclusive Methodology and Technology in Learning and Teaching*, 5(1), 2025.

[15] M. Siino et al. Exploring llms applications in law: A literature review on current legal nlp approaches. *IEEE Access*, 2025.

[16] H. Zhong et al. How does nlp benefit legal system: A summary of legal artificial intelligence. *arXiv:2004.12158*, 2020.

[17] I. Chalkidis et al. Neural legal judgment prediction in english. *arXiv:1906.02059*, 2019.

[18] M. Medvedeva et al. Using machine learning to predict decisions of the european court of human rights. *Artificial Intelligence and Law*, 28(2):237–266, 2020.

[19] N. Aletras et al. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ computer science*, 2:e93, 2016.

[20] G. Garzo et al. Opening the black box: How boolean ai can support legal analysis. In *2024 4th International Conference on Computer Communication and Artificial Intelligence (CCAI)*. IEEE, 2024.

[21] G. Garzo and A. Palumbo. The first hardware circuit emulating italian road homicides legal logic, daje! In *IS 2025-18th IADIS International Conference Information Systems*, 2025.

[22] L. Zheng et al. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 159–168, 2021.

[23] V. Nanda et al. Fairness through robustness: Investigating robustness disparity in deep learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021.

[24] L. Zheng et al. A reasoning-focused legal retrieval benchmark. In *Proceedings of the 2025 Symposium on Computer Science and Law*, pages 169–193, 2025.

[25] I. Chalkidis et al. Lexglue: A benchmark dataset for legal language understanding in english. *arXiv:2110.00976*, 2021.

[26] A. Li et al. Legal judgment prediction based on knowledge-enhanced multi-task and multi-label text classification. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6957–6970, 2025.

[27] J. Rabelo et al. Coliee 2020: methods for legal document retrieval and entailment. In *New Frontiers in Artificial Intelligence: JSAI-isAI 2020 Workshops, JURISIN, LENLS 2020 Workshops, Virtual Event, November 15–17, 2020, Revised Selected Papers 12*, pages 196–210. Springer, 2021.

[28] A. Blair-Stanek et al. Blt: Can large language models handle basic legal text? *arXiv:2311.09693*, 2023.

[29] C. Hurshman. Do opaque algorithms have functions? *Synthese*, 204(3):91, 2024.

[30] R. Zhang et al. Evaluation ethics of llms in legal domain. *arXiv:2403.11152*, 2024.

[31] O. Isaac and O. Ifeoluwa. Artificial intelligence in law: Regulatory adaptation and ethical considerations.

[32] Grazia G. and A. Palumbo. Legal & ethical implications of predictive digital techniques in the judicial criminal proceedings. In *13th International Symposium on Digital Forensics and Security (ISDFS)*, 2025.
[33] G. Garzo and A. Palumbo. Human-in-the-loop: Legal knowledge formalization in attempto controlled english. In *13th International Symposium on Digital Forensics and Security (ISDFS)*, 2025.
[34] L. Floridi. Establishing the rules for building trustworthy ai. *Nature Machine Intelligence*, 1(6), 2019.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251579*

# A Two-Dimensional Evaluation Framework for Factual and Reasoning Assessment of LLMs in Legal Question Answering

Sinan GULTEKIN [a], Matteo ROSSI REICH [a], Francesca GALLONI [a], Francesca LAGIOIA [a,d,1], Elena CONSIGLIO [b,1], Giovanni SARTOR [a,d], Sara BAGNATO [c]

[a] *CIRSFID Alma-AI, Faculty of Law, University of Bologna, Italy*
[b] *University of Palermo, Italy*
[c] *University of Roma LUMSA, Italy*
[d] *European University Institute, Law Department, Italy*

**Abstract.** Deploying Large Language Models (LLMs) for legal question-answering requires ensuring factual accuracy and logical coherence. Current evaluation metrics inadequately capture legal reasoning complexity, while expert assessments lack scalability. We propose a two-dimensional framework that independently measures Truthfulness and Reasoning Soundness in model outputs, applied to Italian asylum proceedings requiring evidence-based analysis. This dual-axis approach reveals critical issues—such as legally correct answers derived through unsound or hallucinatory reasoning—that standard metrics fail to detect. To enable large-scale application, we implement an automated LLM-as-a-Judge system with bias-mitigation techniques. Experimental results demonstrate strong correspondence between automated judgments and expert evaluations, confirming framework reliability. This work advances diagnostic methodology for assessing LLMs in legal domains, offering both theoretical insight and practical applicability toward more trustworthy and accountable legal AI systems.

**Keywords.** LLM, prompting, LLM-as-a-judge, Legal AI

## 1. Introduction

Large Language Models (LLMs) attract growing interest for automating legal question answering (QA), assisting practitioners in extracting information from large volumes of unstructured documents to support scalable decision-making. Yet, the legal field imposes stringent demands on accuracy, explainability, and transparency, which LLMs do not consistently guarantee [1], making their evaluation challenging. Standard automated metrics fail to assess legal correctness or capture reasoning nuances [2], while human

[1]Corresponding Authors: Elena Consiglio at elena.consiglio@unipa.it and Francesca Lagioia at francesca.lagioia@unibo.it

evaluation, though context-sensitive, is resource-intensive, idiosyncratic [3], and prone to declining accuracy under cognitive load [4]. We therefore propose a granular evaluation framework based on the LLM-as-a-Judge paradigm.

Our main contribution is a two-dimensional framework that dissects answers' quality along two dimensions: output accuracy and reasoning soundness. We test such framework on a demanding task, i.e., extracting information from Italian asylum proceedings. This task is representative of real-world practice and well-suited to evaluating the reasoning abilities of LLMs.

This work builds on recent critiques of the LLM-as-a-Judge paradigm. Studies show that even top-performing models exhibit significant shortcomings, especially in areas requiring specialized knowledge [5]. We tackle these limitations by implementing structured prompts, rigorous human-grounded assessments, and a two-dimensional framework applied to information extraction from a dataset of Italian asylum cases. By focusing on legal context, we provide a scalable and reliable evaluation methodology applicable in high-stakes domains, where expert knowledge is essential and extractive QA alone is insufficient to answer a question.

The paper is structured as follows: Section 2 reviews the related works, Section 3 describes the dataset and Section 4 presents the two-dimensional evaluation framework and its implementation. Section 5 reports the experimental results. Finally, Section 6concludes and outlines directions for future research.

## 2. Related Works

The evaluation of natural language generation (NLG) has significantly evolved, moving from simple lexical comparisons to sophisticated, reasoning-based, assessments. This reflects both the increasing complexity of NLG tasks and the growing capabilities of models. Early automated metrics like BLEU[6], METEOR[7], and ROUGE [8] measured surface-level similarity through n-gram analysis. Though scalable and consistent, they struggle with semantic equivalence and penalize valid paraphrasing, producing weak correlation with human judgments [9].

Subsequent methods incorporated semantic understanding through contextual embeddings. BERTScore [10] enabled synonym recognition via token-level similarity, while BLEURT [11] and COMET [12] improved correlations with human assessments through fine-tuning on rating data. However, these methods may not capture higher-order qualities like logical coherence, factual grounding, or argument soundness.

The emergence of powerful LLMs enables a new paradigm, i.e., models acting as evaluators, a concept commonly known as "LLM-as-a-Judge" [13]. Leveraging advanced instruction-following and reasoning capabilities, LLMs emulate context-sensitive expert analysis. In recent studies, LLMs achieve high agreement with human judgments across tasks, ranging from ranking chatbot responses to evaluating output safety and helpfulness, all at greater scalability and lower cost [13,14].

Recent research emphasizes granular diagnostic evaluations. FActScore [15] pioneered atomic evaluation by verifying single factual claims against knowledge sources, while ARES [16] assessed RAG output faithfulness and contextual relevance. G-Eval [17] employed Chain-of-Thought prompting for step-by-step logic analysis, enhancing transparency, alongside methods targeting logical fallacies [18].

The evalution of LLMs' outputs is critical in the legal domain, which demands impeccable factual and legal accuracy, as well as transparent and verifiable reasoning [19]. Legal NLP made substantial progress in creating domain-specific benchmarks like LegalBench [20], which enables systematically measures various facets of legal reasoning. Yet evaluations still reveal limits. Even state-of-the-art LLMs show instability in answering legal questions [21] and the analyses often remain brief and unreliable, even where they follow basic structures like IRAC (Issue, Rule, Analysis, Conclusion), [22].

Most critically, hallucinations remain widespread [23], with rates between 59% and 88% for some queries, casting doubts on the readiness of LLMs for legal practice. While fine-tuning methods are proposed to improve accuracy [24], they usually address either grounding or reasoning in isolation.

Our contribution introduces a two-dimensional evaluation framework that concurrently assesses both truthfulness and reasoning soundness. We operationalize this framework through a specialized LLM-as-a-Judge system, tailored to the evidence-based analysis required in the legal field, providing a granular and diagnostic tool for assessing LLMs' reliability in high-stakes areas, where errors may carry serious consequences.

## 3. Dataset

As mentioned, in this study, we focus on applications for international protection in Italy. The dataset consists of legal and administrative texts from 91 asylum cases appealed to the Tribunal of Palermo and decided between 2014 and 2023. Of these, asylum was granted in 19 cases, humanitarian protection in 31, subsidiary protection in 21, and 20 cases were fully rejected. All personally identifiable information has been removed.

The procedure begins with an asylum claim, followed by identification and completion of the C3 Form questionnaire. The Territorial Commission evaluates claims and may: 1) recognize refugee status or subsidiary protection; 2) forward paperwork for humanitarian/special protection; or 3) reject the claim. Rejected or unsatisfactory decisions may be appealed to the judicial authority.

The dataset[2] contains decisions appealed to the Tribunal of Palermo, that is competent to adjudicate decisions of Territorial Commissions in Western Sicily (Palermo, Trapani, Agrigento). Representativeness is limited: the dataset contains only decisions from Western Sicily Commissions that were rejected or unsatisfactory, as authorities denied broader access and only cooperating law firms provided cases under appeal. Additionally, most cases involve male applicants, reflecting broader refugee flow patterns but limiting female asylum seeker analysis. The documentation for each case file is composed of one or more of the following documents:

- C3 Form (Modulo C3): A standardised questionnaire detailing the applicant's personal and family information, normally compiled by the police;
- Interview Transcript (Audizione): A verbatim record of the personal interview with the staff of the Territorial Commission, during which the applicant states the grounds for their claim, often conducted with the help of an interpreter. The interview might be recorded and the transcript might be obtained, using AI, from a video recording;

[2] Available on GitHub at https://github.com/EquitableAlgorithms/Two-Dimensional-Evaluation-Framework

- Initial Decision by the Territorial Commission (Commissione): The official ruling, stating the determination on the asylum claim and detailing the reasons for the decision.
- First-Tier Tribunal's Judgment (Tribunale): The judgment from the court of first instance (Tribunale), confirming or revising the initial determination by the administrative authority, with reasons.

Starting from PDF scans, we process these files using Surya [25], a transformer-based OCR toolkit, to create a machine-readable corpus. Given the highly sensitive and personal nature of the contained information, we apply a rigorous anonymisation protocol, to remove all Personally Identifiable Information (PII).

## 4. Methodology

Our methodology is designed to create a robust, scalable and fine-grained evaluation of LLMs performance in the legal area. As shown in Figure 1, the workflow includes the following steps.

The first step is *Questions and Guidelines Design*.[3] After examining a large set of cases, for each type of document, we identified relevant questions. To avoid biases, we adopted a bottom-up approach: questions were meant to extract any "interesting" information, regardless of its legal relevance to the decision of the case. Additionally, we drafted a set of guidelines to be used by legal experts in answering the same questions. Both guidelines and questions were iteratively refined based on experts' feedback.

The second step is *Ground-truth Construction*. For each case, four legal experts independently answered all questions.

The third step is the *Information Extraction*. For each case, LLMs were asked to address the mentioned questions, providing two outputs: (a) the answer and (b) the reasoning justifying such answer, grounded in the text of the decision.

The fourth step is the *Answer Evaluation*. Both, experts and LLM-judges, evaluated the LLMs answers.

The last step is *Performance comparison.* We assessed the LLM-judges evaluations against those of experts, to determine the former's performance.

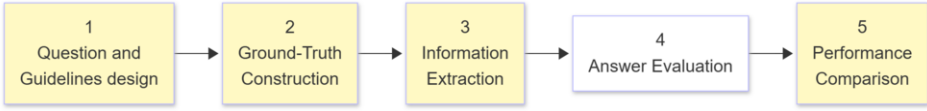



**Figure 1.** Evaluation Pipeline

### *4.1. The Methodology for Evaluating the LLMs Output*

The evaluation protocol aims to ensure a consistent and nuanced evaluation of LLMs-generated answers[4]. It proceeds along two primary dimensions: *Truthfulness* and *Reasoning Soundness*, as shown in Figure 2.

[3]The question set is available on GitHub at https://github.com/EquitableAlgorithms/Two-Dimensional-Evaluation-Framework

[4]The evaluation guidelines are available on GitHub at https://github.com/EquitableAlgorithms/Two-Dimensional-Evaluation-Framework

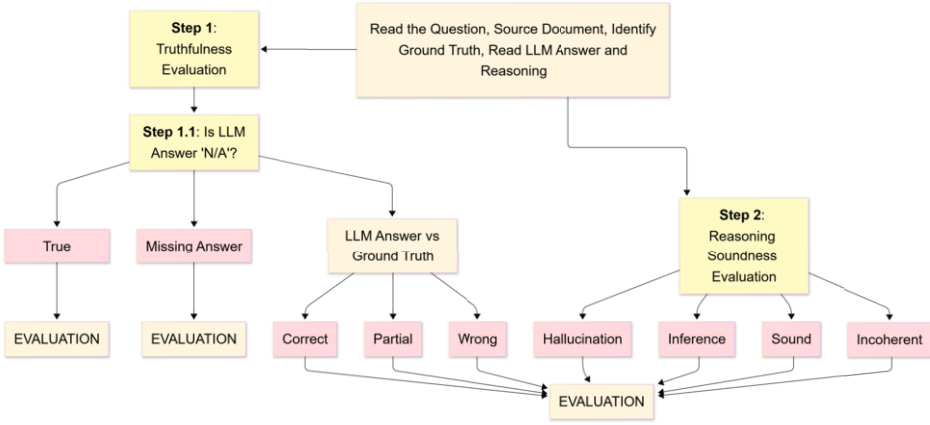



**Figure 2.** Evaluation Workflow

The *Truthfulness Evaluation* assesses the answer by LLMs, distinguishing between 'N/A' (information is Not Available in the text) answers, and those providing a definite response to a YES/NO or an open question. 'N/A' answers are labelled as follows: '*True Negative*' if the requested information is absent from the document; or '*Missing Answer*' if it is present. Definite answers are compared against the Ground Truth, and labelled as follows: '*Correct*' if it accurately matches the Ground Truth, regardless of minor phrasal differences; '*Partial*' if it is incomplete or mixes correct with irrelevant or incorrect information; '*Wrong*' if it contradicts the Ground Truth or provides entirely unrelated information.

The *Reasoning Soundness Evaluation* labels the LLMs' *reasoning* as follows: '*Incoherent*' if it is nonsensical or completely off-topic; '*Hallucination*' if it incorporates invented information, absent in the source document; '*Inference*' if it consists in inferences, leading to conclusion that are not necessarily entailed by the text (while being consistent with such a text and even somehow supported by it): '*Sound*' if it is based on the text, with a clear understanding of the question and without fabrication.

The combination of Truthfulness and Reasoning Soundness provides an analytical assessment of the quality and reliability of information extraction, as shown in Table 1.

A crucial aspect of our methodology pertains to ensuring the correctness and consistency of the ground truth. Thus, legal experts proceed according to the following steps: (i) thoroughly understand the question, (ii) review the source, (iii) establish the Ground Truth, and (iv) examine the LLMs' answers and their reasoning.

After evaluating an initial set of cases, experts were given an alignment batch (939 question–answering and reasoning items) to compute the inter-evaluation agreement and solve doubts. Given the open-ended formulation of questions, the complexity of documents and the diversity of contents, with many inherently ambiguous borderline cases, the agreement rate of 75.7% was satisfactory. Thus, we proceeded with the distribution of the remaining cases. Disagreement often arose because several points could be pertinent in a single case, leading to context-dependent interpretations and multiple justifiable reasoning paths. This complexity reflects the richness of legal evaluation rather than inconsistency among annotators, and it highlights the challenge of defining a single, unambiguous ground truth in such domains. Thus, we proceeded with the distribution of the remaining cases

**Table 1.** Legal Experts Evaluation

| Questions | LLM Answer | LLM Reasoning | Evaluation | Rationale | Ground Truth |
|---|---|---|---|---|---|
| Q3: What relationship did the applicant have with their family members? What relationship do they have now with their family members? | In the past, they lived with their father, mother, and three sisters. Currently, they have lost contact with the mother, and there is no information about the sisters. | The text states: 'To have lived with my family, consisting of my father and mother. I have three younger sisters.' and 'To have no more news of my mother since the day she was taken by the rebels'. | Partial, Hallucination | Information about the sisters is not present in the text. | They lived with the father and mother; the current relationship is not mentioned. |
| Q3.1: Does the commission consider the elements related to problems with family members to be relevant? (YES, NO, N/A) | NO | The text does not indicate that the commission considers the elements related to problems with family members to be relevant. The commission focuses on other aspects of the applicant's case. | Wrong, Inference | The commission does not dwell on these aspects; the question in question cannot be answered. | N/A |

### *4.2. LLMs-Based Information Extraction*

For the information extraction, we employ a "Few-shot-CoT" approach [26,27], including 5 examples, 2 YES/NO questions, and 3 open questions. Examples were selected so as to direct the LLM to answer the most difficult questions correctly. We run experiments with smaller and bigger models; the latter enabled a considerable jump in F1score on the information extraction task, as can be seen from Table 2. Therefore, we experiment with 2 prompts on the big models [5] to understand the performance on the downstream task of automatic answer evaluation. The first prompt instructs the LLM to impersonate three debating lawyers who must reach a majority consensus. The second one implements a *disputatio*-style debate, wherein one persona argues for a proposed answer while a second contends that the answer is "N/A". Both prompts, optimized using an LLM [28], ask for a structured JSON output and require the model to provide its *reasoning* for each answer.

As reported in Table 3, the first prompt had a marginally higher Truthfulness Accuracy (89.81% vs. 87.92%). The second prompt reached a significantly higher Reasoning Soundness Accuracy (78.69% vs. 75.83%). Moreover, it demonstrated slightly better overall answer quality, with higher scores in both Exact Match Ratio (70.00% vs. 68.75%) and F1 Score (0.94 vs. 0.93).

We selected the second prompt considering that its superiority in reasoning was crucial for performance evaluation in the LLM-as-a-judge framework.

**Table 2.** Information Extraction performance

(a) Performance with P.1

| LLM QA Model | Accuracy | F1 Score |
|---|---|---|
| Mistal 7B | 65.3% | 0.56 |
| Qwen 2.5 7B | 59.3% | 0.55 |
| Overall | 62.3% | 0.55 |
| Gemma 3 27B | 72.1% | 0.71 |
| o4-mini | **82.4%** | **0.80** |
| Overall | 77.3% | 0.75 |

(b) Performance with P.2

| LLM QA Model | Accuracy | F1 Score |
|---|---|---|
| Mistal 7B | 66.2% | 0.57 |
| Qwen 2.5 7B | 65.7% | 0.56 |
| Overall | 65.9% | 0.57 |
| Gemma 3 27B | 69.5% | 0.66 |
| o4-mini | **81.8%** | **0.79** |
| Overall | 75.7% | 0.73 |

[5] The feature extraction prompts are available on GitHub at https://github.com/EquitableAlgorithms/Two-Dimensional-Evaluation-Framework

**Table 3.** Prompt techniques for Information Extraction

| QA Model | EM Ratio | Overall F1 | Truth. Acc. | Truth. F1 | Reason. Acc. | Reason. F1 | Avg. Time |
|---|---|---|---|---|---|---|---|
| G. 27b P.1 | 63.33% | 0.92 | 88.78% | 0.89 | 71.09% | 0.71 | 22'43" |
| o4-mini P.1 | 74.17% | 0.94 | 90.83% | 0.91 | 80.58% | 0.81 | 22'43" |
| P.2 Overall | 68.75% | 0.94 | **89.81%** | **0.90** | 75.83% | 0.76 | 22'43" |
| G. 27b P.2 | 65.64% | 0.93 | 87.95% | 0.88 | 73.85% | 0.74 | 15'17" |
| o4-mini P.2 | 74.36% | 0.95 | 87.88% | 0.89 | 83.53% | 0.84 | 28'46" |
| P.2 Overall | **70.00%** | **0.94** | 87.92% | 0.88 | **78.69%** | **0.79** | **22'02"** |

We sent the questions to the models in batches,increase efficiency. An automated validation layer then parsed the outputs to ensure that every question was answered correctly. If a question was missed, a targeted re-prompting mechanism queried the model with the specific unanswered question up to five times, significantly enhancing the system's robustness and overall reliability.

### *4.3. LLM-as-a-Judge*

To enable evaluation at scale, we developed an "LLM-as-a-Judge" framework [13]. This approach combines scalability with explanatory reasoning, to provide pointwise evaluation of content accuracy, including a final rating and a descriptive feedback.

The prompt for LLM-as-a-Judge includeed our detailed evaluation guidelines to direct the evaluation [6].

A critical challenge in deploying LLMs as evaluators is their susceptibility to biases that can compromise reliability. Our framework incorporated several design choices to mitigate such biases. We addressed presentation-related biases such as **positional** [13,29] and **verbosity** bias [13,30] by employing a pointwise, reference-based evaluation: the LLM-judge assessed each single candidate answer against the corresponding ground truth. This eliminated the structural opportunity for the model to prefer one answer over another based on presentation order or length. We addressed content and cognitive biases through prompt engineering. To counter **authority** bias [30], where seemingly authoritative claims might sway a model, our prompt explicitly instructed the judge to verify all answers against the provided source document. This forced a grounded evaluation, preventing the model from accepting unsubstantiated assertions. Furthermore, the disputation mechanism, which simulates an adversarial debate, reduced **overconfidence** bias by forcing the judge to consider counterarguments, a technique shown to improve truthfulness [31]. Finally, to prevent **self-enhancement** bias, where a model tends to rate its outputs more favorably [32,33], the judge's model was different from the one producing the output being evaluated. This architectural separation ensures the evaluator has no vested interest in the generated output, promoting a more objective assessment.

According to our guidelines, the evaluation included distinct "explanation-rating" steps to enhance the reliability of the final rating. The pipeline consisted of the following steps. First the LLM checks whether the evaluated model has refused to answer, returning 'N/A'. For YES/NO or open answers, the system proceeds to a two-step evaluation. First, it assesses the Truthfulness of the answer by comparing it against the ground truth,

[6]The LLM-as-a-judge prompt is available on GitHub at https://github.com/EquitableAlgorithms/Two-Dimensional-Evaluation-Framework

identifying discrepancies, and rendering a verdict. Then it assesses the Reasoning Soundness by matching the LLM's reasoning with both the human-provided reasoning and the whole source documents. This technique facilitates identifying hallucinations [17]. This structured, step-by-step process also enhances the interpretability of the judge's decision-making, ensuring more transparent evaluations.

## 5. Experimental Results

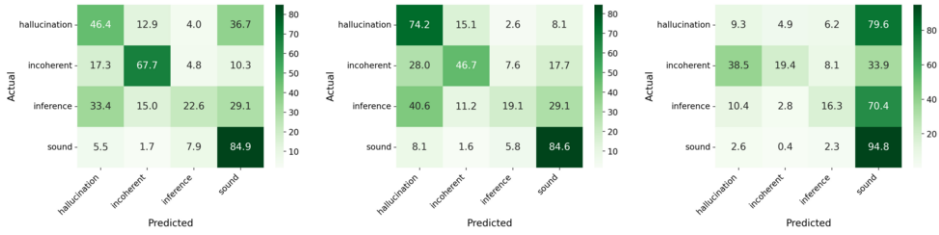
Gemini 2.5 Flash (G2.5 F.) was evaluated on the full dataset as baseline. We compared Gemini 2.5 Flash Lite (G2.5 F. L.), Gpt-5 Nano, Mistral Small 24B 2501, and Qwen3 14B on a stratified subset of 20 cases (each ¿300 questions), enabling the 'Thinking' feature for energy-efficient comparison. The results of this comparative evaluation are presented in Table 4 and Figure 3.

**Table 4.** A comparison of Judge model performance metrics

| Judge Model | EM Ratio | Truth. Accuracy | Truth. F1 Score | Reasoning Accuracy | Reasoning F1 Score | Avg. Time per Case |
|---|---|---|---|---|---|---|
| Gem. 2.5 F. Full | 71.3% | 90.4% | 0.92 | 77.8% | 0.80 | **26'45"** |
| Gem. 2.5 F. | 71.4% | **90.5%** | **0.92** | 77.8% | 0.80 | **26'45"** |
| Gem. 2.5 F. L. | 70.3% | 89.6% | 0.92 | 77.0% | 0.80 | 142'48" |
| Gpt-5 Nano | **74.5%** | 89.8% | 0.92 | **82.6%** | **0.81** | 132'31" |
| Mistral S. 24B | 53.8% | 88.4% | 0.90 | 56.0% | 0.65 | 32'20" |
| Qwen 3 14B | 69.9% | 87.5% | 0.90 | 77.5% | 0.79 | 58'15" |

The most immediate observation is the substantial difference in computational cost; enabling the 'thinking' feature increased the average processing time per case up to five-fold, rendering the reasoning approach less practical for scalable deployment. Moreover, all models consistently yielded higher accuracy in Truthfulness than Reasoning Soundness, suggesting that answering is considerably less complex objective than reasoning. While GPT-5 Nano (Thinking) initially appeared to achieve the highest Reasoning Accuracy (82.6%), a deeper analysis of confusion matrices revealed that this is misleading. The model exhibited a strong bias toward the majority class ('sound'), while performing poorly on minority classes (e.g., 'hallucinations'), correctly identifying only 6.2%. In contrast, the Gemini 2.5 Flash models demonstrated a more consistent performance across categories: Gemini 2.5 Flash Light (Thinking) particularly excelled in spotting hallucinations, indicating a more robust, less biased evaluation capability, while the standard Gemini 2.5 Flash provided a superior balance of speed and low computational cost, emerging as the most viable option. On the smaller models side, Mistral-small was outperformed on overall metrics by the rest of its competitors; however, it showed a more balanced performance across the different classes. Even though Qwen3-14B seemed a small, promising alternative to others, a close examination of the confusion matrix and computational time shows that Gemini models predicted classes other than Sound more reliably. A qualitative review of the evaluation process highlighted two significant and recurring challenges. First the LLM Judge generally struggles in differentiating between "inference" and "hallucinations", this happens when the model has to decide if a conclusion is logically derived from premises within the text or if it's new information be-

ing incorrectly presented as fact. Second, the consistent application of the 'Partial' label proved difficult due to its inherent subjectivity. The guideline defines this based on "missing crucial information", which forces a value judgment on the part of the human annotator.



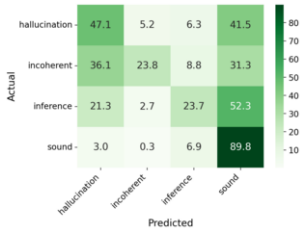

(a) Gemini 2.5 Flash



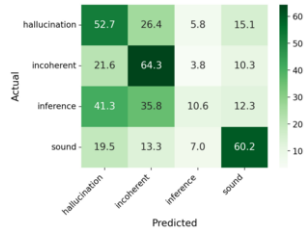

(b) Gemini 2.5 Flash Light



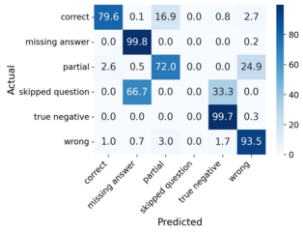

(c) GPT-5 Nano



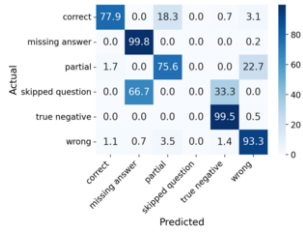

(d) Qwen 3 14B



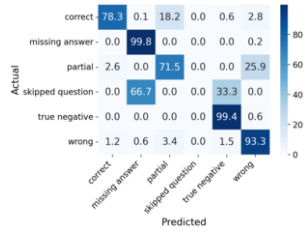

(e) Mistral Small 24B



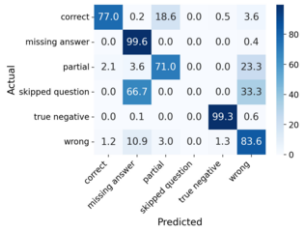

(f) Gemini 2.5 Flash



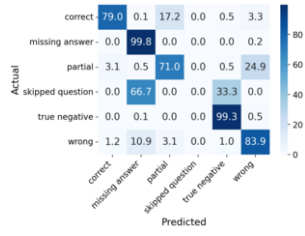

(g) Gemini 2.5 Flash Light




(h) GPT-5 Nano




(i) Qwen 3 14B




(j) Mistral Small 24B

**Figure 3.** Confusion matrices for all judge models. **Rows 1-2:** Reasoning Soundness evaluation. **Rows 3-4:** Truthfulness evaluation.

### 5.1. Computational Cost

Local models were benchmarked on a server equipped with a `Tesla V100-PCIE-32GB` GPU. The inference time per question was influenced by the number of retrials required after a failure to answer, an issue strongly emerging when using the second, more complex prompt. For the first prompt, the `mistralai/Mistral-7B-Instruct-v0.3` and `Qwen/Qwen2.5-7B-Instruct` models recorded inference times of 28.26s and 21.73s, respectively. For the second prompt, these times increased significantly to 90.38s and 150.96s. In contrast, API-based models showed a more contained increase: `google/gemma-3-27b-it` went from 17.84s to 25.76s, and `o4-mini-2025-04-16` remained at 8.26s for both prompts, reflecting their lower retrial rates. The LLM judge, using `gemini-2.5-flash-preview-05-20`, averaged 5.13 seconds per question. Budget is a critical constraint, as each case averages 7.9M input and 3.8M output tokens, making the per-token price of flagship models like Gemini 2.5 Pro prohibitively expensive.

## 6. Conclusion

We introduced a novel, two-dimensional framework for the granular evaluation of LLMs in legal QA. By distinguishing performance in Truthfulness and Reasoning Soundness, we obtained a more nuanced assessment. We operationalized this framework using an LLM-as-a-Judge system, incorporating adversarial prompting and architectural choices to mitigate evaluative biases. Our experiments, grounded in a real-world dataset of Italian asylum cases, demonstrate the viability of our approach.

Our findings indicate that while current models can assess truthfulness with high accuracy, the evaluation of reasoning soundness remains a significant challenge. Results revealed a performance gap across all tested models, highlighting the complexity of scrutinizing logical chains of thought.

There is a clear need to refine the LLM-as-a-Judge's capacity for evaluating reasoning, potentially through more sophisticated prompting techniques or model fine-tuning for logical analysis. Exploring the framework's applicability to other complex legal tasks, such as case-brief generation or statutory interpretation, is a further promising avenue for extending this research and moving towards more reliable and transparent AI applications in the legal domain.

## 7. Acknowledgments

This work was partially supported by the following projects: CompuLaw – Computable Law – funded by the ERC under the Horizon 2020 (Grant Agreement N. 833647); PRIN2022 PRIMA - PRivacy Infringements Machine-Advice (Ref. Prot. n.: 20224TPEYC - CUP J53D23005130001); PRIN2022 EQUAL – EQUitableALgorithms (Ref. Prot n. 2022KFLF3E_001 - CUP J53D23005560001); "FAIR - Future Artificial Intelligence Research" – Spoke 8 and Spoke 1 (CAI4DSA action) under the European Commission's NextGeneration EU programme, PNRR – M4C2 – Inv. 1.3, Partenariato Esteso (PE00000013)

## References


[1] Doyle C, Tucker AD. If You Give an LLM a Legal Practice Guide. In: Proceedings of the 2025 Symposium on Computer Science and Law; 2025. p. 194-205.

[2] Schluter N. The limits of automatic summarisation according to ROUGE. In: Lapata M, Blunsom P, Koller A, editors. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. Valencia, Spain: Association for Computational Linguistics; 2017. p. 41-5. Available from: https://aclanthology.org/E17-2007/.

[3] Clark E, August T, Serrano S, Haduong N, Gururangan S, Smith NA. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. In: Zong C, Xia F, Li W, Navigli R, editors. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics; 2021. p. 7282-96. Available from: https://aclanthology.org/2021.acl-long.565/.

[4] Kern C, Eckman S, Beck J, Chew R, Ma B, Kreuter F. Annotation Sensitivity: Training Data Collection Methods Affect Model Performance. In: Bouamor H, Pino J, Bali K, editors. Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: Association for Computational Linguistics; 2023. p. 14874-86. Available from: https://aclanthology.org/2023.findings-emnlp.992/.

[5] Szymanski A, Ziems N, Eicher-Miller HA, Li TJJ, Jiang M, Metoyer RA. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. In: Proceedings of the 30th International Conference on Intelligent User Interfaces; 2025. p. 952-66.

[6] Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a Method for Automatic Evaluation of Machine Translation. In: Isabelle P, Charniak E, Lin D, editors. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics; 2002. p. 311-8. Available from: https://aclanthology.org/P02-1040/.

[7] Rei R, Stewart C, Farinha AC, Lavie A. COMET: A Neural Framework for MT Evaluation. In: Webber B, Cohn T, He Y, Liu Y, editors. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics; 2020. p. 2685-702. Available from: https://aclanthology.org/2020.emnlp-main.213/.

[8] Lin CY. ROUGE: A Package for Automatic Evaluation of Summaries. In: Text Summarization Branches Out. Barcelona, Spain: Association for Computational Linguistics; 2004. p. 74-81. Available from: https://aclanthology.org/W04-1013/.

[9] Reiter E. A Structured Review of the Validity of BLEU. Computational Linguistics. 2018 Sep;44(3):393-401. Available from: https://aclanthology.org/J18-3002/.

[10] Zhang* T, Kishore* V, Wu* F, Weinberger KQ, Artzi Y. BERTScore: Evaluating Text Generation with BERT. In: International Conference on Learning Representations; 2020. Available from: https://openreview.net/forum?id=SkeHuCVFDr.

[11] Sellam T, Das D, Parikh A. BLEURT: Learning Robust Metrics for Text Generation. In: Jurafsky D, Chai J, Schluter N, Tetreault J, editors. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. p. 7881-92. Available from: https://aclanthology.org/2020.acl-main.704/.

[12] Rei R, Stewart C, Farinha AC, Lavie A. COMET: A Neural Framework for MT Evaluation. In: Webber B, Cohn T, He Y, Liu Y, editors. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics; 2020. p. 2685-702. Available from: https://aclanthology.org/2020.emnlp-main.213/.

[13] Zheng L, Chiang WL, Sheng Y, Zhuang S, Wu Z, Zhuang Y, et al. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23. Red Hook, NY, USA: Curran Associates Inc.; 2023. .

[14] Chan CM, Chen W, Su Y, Yu J, Xue W, Zhang S, et al. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. ArXiv. 2023;abs/2308.07201:null.

[15] Min S, Krishna K, Lyu X, Lewis M, Yih Wt, Koh P, et al. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In: Bouamor H, Pino J, Bali K, editors. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics; 2023. p. 12076-100. Available from: https://aclanthology.org/2023.emnlp-main.741/.

[16] Saad-Falcon J, Khattab O, Potts C, Zaharia M. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. In: Duh K, Gomez H, Bethard S, editors. Proceedings of

the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Mexico City, Mexico: Association for Computational Linguistics; 2024. p. 338-54. Available from: https://aclanthology.org/2024.naacl-long.20/.

[17] Liu Y, Iter D, Xu Y, Wang S, Xu R, Zhu C. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In: Bouamor H, Pino J, Bali K, editors. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics; 2023. p. 2511-22. Available from: https://aclanthology.org/2023.emnlp-main.153/.

[18] Cheng Q, Sun T, Zhang W, Wang S, Liu X, Zhang M, et al.. Evaluating Hallucinations in Chinese Large Language Models; 2023. Available from: https://arxiv.org/abs/2310.03368.

[19] Zhong M, Zhang A, Wang X, Hou R, Xiong W, Zhu C, et al.. Law of the Weakest Link: Cross Capabilities of Large Language Models; 2024. Available from: https://arxiv.org/abs/2409.19951.

[20] Guha N, Nyarko J, Ho DE, Ré C, Chilton A, Narayana A, et al. LEGALBENCH: a collaboratively built benchmark for measuring legal reasoning in large language models. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23. Red Hook, NY, USA: Curran Associates Inc.; 2023. .

[21] Blair-Stanek A, Durme BV. LLMs Provide Unstable Answers to Legal Questions. ArXiv. 2025;abs/2502.05196. Available from: https://api.semanticscholar.org/CorpusID:276249373.

[22] Peoples L. Artificial Intelligence and Legal Analysis: Implications for Legal Education and the Profession; 2025. Available from: https://arxiv.org/abs/2502.03487.

[23] Dahl M, Magesh V, Suzgun M, Ho DE. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. Journal of Legal Analysis. 2024 06;16(1):64-93. Available from: https://doi.org/10.1093/jla/laae003.

[24] Hu Y, Gan L, Xiao W, Kuang K, Wu F. Fine-tuning Large Language Models for Improving Factuality in Legal Question Answering. In: Rambow O, Wanner L, Apidianaki M, Al-Khalifa H, Eugenio BD, Schockaert S, editors. Proceedings of the 31st International Conference on Computational Linguistics. Abu Dhabi, UAE: Association for Computational Linguistics; 2025. p. 4410-27. Available from: https://aclanthology.org/2025.coling-main.298/.

[25] Paruchuri V, Team D. Surya: A lightweight document OCR and analysis toolkit; 2025. GitHub repository. https://github.com/VikParuchuri/surya.

[26] Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22. Red Hook, NY, USA: Curran Associates Inc.; 2022. .

[27] Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22. Red Hook, NY, USA: Curran Associates Inc.; 2022. .

[28] Gonen H, Iyer S, Blevins T, Smith N, Zettlemoyer L. Demystifying Prompts in Language Models via Perplexity Estimation. In: Bouamor H, Pino J, Bali K, editors. Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: Association for Computational Linguistics; 2023. p. 10136-48. Available from: https://aclanthology.org/2023.findings-emnlp.679/.

[29] Shi L, Ma C, Liang W, Diao X, Ma W, Vosoughi S. Judging the Judges: A Systematic Study of Position Bias in LLM-as-a-Judge; 2025. Available from: https://arxiv.org/abs/2406.07791.

[30] Ye J, Wang Y, Huang Y, Chen D, Zhang Q, Moniz N, et al.. Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge; 2024. Available from: https://arxiv.org/abs/2410.02736.

[31] Khan A, Hughes J, Valentine D, Ruis L, Sachan K, Radhakrishnan A, et al.. Debating with More Persuasive LLMs Leads to More Truthful Answers; 2024. Available from: https://arxiv.org/abs/2402.06782.

[32] Panickssery A, Bowman SR, Feng S. LLM Evaluators Recognize and Favor Their Own Generations; 2024. Available from: https://arxiv.org/abs/2404.13076.

[33] Li R, Patel T, Du X. PRD: Peer Rank and Discussion Improve Large Language Model based Evaluations; 2024. Available from: https://arxiv.org/abs/2307.02762.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251580*

# Can LLMs Create Legally Relevant Summaries and Analyses of Videos?

Lyra HOEBEN-KUIL [a,1], Gijs VAN DIJCK [a,b], Jaromir SAVELKA [c], Johanna GUNAWAN [a], Konrad KOLLNIG [a], Marta KOŁACZ [a], Mindy DUFFOURC [a], Shashank M CHAKRAVARTHY [b], Hannes WESTERMANN [a]

[a] *Maastricht Law and Tech Lab, Faculty of Law, Maastricht University, The Netherlands*
[b] *Brightlands Institute for Smart Society, Maastricht University, The Netherlands*
[c] *Computer Science Department, Carnegie Mellon University, Pittsburg, USA*

**Abstract.** Understanding the legally relevant factual basis of an event and conveying it through text is a key skill of legal professionals. This skill is important for preparing forms (e.g., insurance claims) or other legal documents (e.g., court claims), but often presents a challenge for laypeople. Current AI approaches aim to bridge this gap, but mostly rely on the user to articulate what has happened in text, which may be challenging for many. Here, we investigate the capability of large language models (LLMs) to understand and summarize events occurring in videos. We ask an LLM to summarize and draft legal letters, based on 120 YouTube videos showing legal issues in various domains. Overall, 71.7% of the summaries were rated as of high or medium quality, which is a promising result, opening the door to a number of applications in e.g. access to justice.

**Keywords.** Multi-modal Large Language Models, Generative AI, Law & AI, Access to Justice, Video Analysis

## Introduction

Many people lack adequate access to justice, leaving them unable to resolve their legal problems [6], with significant costs to both individuals and society [27]. A central difficulty lies in the translation of lived experience into (legally) relevant text narratives [39]. Laypeople often have difficulties identifying which facts are legally relevant and how to put the facts into a legally valid textual representation [3,32]. Whether in insurance claims, court submissions, or landlord–tenant and consumer disputes, individuals must present facts in a structured, legally meaningful way. The difficulty of people in crafting narratives from their experiences can present an important barrier in obtaining access to justice and in properly utilizing governmental resources [23,3,21].

Researchers have explored the use of AI to bridge the gap between a layperson understanding of events and a legally relevant explanation [39,10,30,4]. However, these approaches still rely on the user interacting with the AI system through text. This means

[1] Corresponding Author: Lyra Hoeben-Kuil, Maastricht University, lyra.hoeben-kuil@maastrichtuniversity.nl

that laypeople who do not have the linguistic skills or vocabulary to describe their issue or identify the relevant details may be left behind.

In this paper, we explore if multi-modal large language models (LLMs) could analyze and derive legally relevant summaries and legal implications from *videos*. Legal relevancy here refers to the extent to which information, facts, or observations contained in the source (a video) contribute to identifying, interpreting, or evaluating issues that have legal implications. If LLMs have this capacity, it could enable laypeople to simply record a video of their legal issue (such as deficiencies in rented apartments, traffic incidents or product malfunctions) and then use an LLM to fill out a form or draft a legal letter. Thus, the barrier to receive legal help for their situation from AI can be further reduced.

To investigate this capability of LLMs, we collect a dataset of videos showing a legal issue from YouTube, and set up an evaluation framework to explore the capacity of the models to process the data. We aim to answer the following research questions:

RQ1 Can LLMs reliably summarize legally relevant facts from video footage?
RQ2 How do video characteristics (e.g., domain, perspective, voice-over, text, and complexity) influence the quality of LLM outputs?
RQ3 To what extent can LLMs link these facts to the relevant legal rules (i.e. the legal provisions that apply to the matter at hand)?

The contributions of this paper are thus as follows:

1. As far as we are aware, the first study exploring the legal analysis of videos using LLMs.
2. A dataset containing a collection of video snippet showing legal issues, along with metadata such as domain, complexity and human-written summaries.
3. A quantitative and qualitative analysis of the capacity of LLMs in summarizing videos and linking them to relevant laws, as well as the factors affecting the quality of the generated text.

## 1. Related Work

LLMs have been actively explored in the task of extracting information from video [13,28]. Alayrac et al. introduced Flamingo—an approach where the model would accept sequences of video frames and perform open-ended tasks such as question answering and captioning [1]. Zhang et al. introduced Video-LLaMA for video comprehension through identification of temporal change in visual scenes [44]. Maaz et al. introduced Video-ChatGPT which is a form of video-based conversation that merges a video-adapted visual encoder with an LLM trained on a diverse dataset comprising of 100,000 video-instruction pairs [20]. Lee et al. proposed a novel video summarization framework, called LLM-based video Summarization (LLMVS) which translated video frames into sequences of captions using a Multi-modal LLM [18]. Earlier works in video summarization, focusing on courtroom video recordings, utilized unsupervised learning by extracting semantic information through automatic speech transcriptions, automatic audio and video annotations [9].

In the legal domain, researchers have extensively explored LLMs' abilities in several key areas such as legal reasoning [42,17,16,2,22], or document analysis and annotation [25,12,26]. Similarly, access to justice initiatives have evolved significantly from the

work such as Protection Order Advisory system [5], web-based legal decision support systems [43] or more recent approaches like the JusticeBot methodology [37,35], combining case-based and rule-based reasoning to help laypeople explore their legal rights. LLMs have also shown promise in dispute mediation [41,34] and providing legal information to laypeople [33]. Recent research has demonstrated their capability to transform legal articles into structured representations for legal decision support tools [14] and create automated forms [29], as well as facilitating the assessment of eligibility for legal aid [30]. LLMs have also been successfully applied to discover common patterns in fact descriptions of court opinion documents [7,11]. A promising use-case for LLMs in access to justice lies in assessing layperson narratives [10,39] or layperson provided-pictures [40] to help them in understanding their legal situation or filling in forms [36]. Here, we expand on this research to study if LLMs can enable laypeople to provide information for such purposes via video clips.

## 2. Proposed Framework

We suggest a framework that uses an LLM to analyze the legally relevant aspects of that video. First, we ask the LLM to draft a summary of what happens in the video, featuring the legally essential factual occurrences. Then, we ask the model to draft a legal complaint letter, which refers to Dutch legislation and states which norms were violated and how.

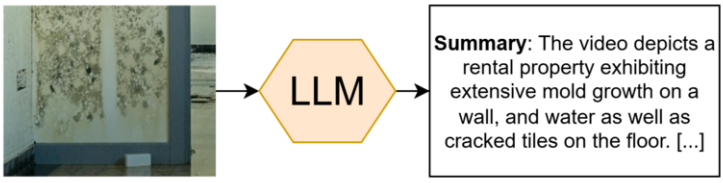



**Figure 1.** Visualization of the pipeline. Input is a video, output is a summary or a legal complaint letter (not shown).

Such a pipeline could have various practical applications, including for access to justice. Laypeople struggle to know what to do when facing legal issues, or how to articulate what has happened in a legally coherent manner, which can present an important barrier to access to justice, as highlighted by several studies [23,3,31,6,24]. If LLMs can summarize and analyze video clips, they could help bridge the gap between layperson user and legal support. An app could ask a user to film the problem they face, and then either provide legal information (coupled with a system such as JusticeBot [38]) or even take the relevant information and fill out relevant forms or draft letters. Even if not perfect, guiding people to a relevant resource or having a draft that can be reviewed with a lawyer could already be a significant improvement.

Even beyond access to justice, such a system could be very powerful in various AI & Law applications. Videos represent one of the most abundant and high-fidelity datasources available to us, and are frequently used as, e.g., evidence. However, until recently, they have mostly been beyond the scope of legal AI systems. We set up experiments to understand the current capacity of LLMs on understanding videos.

## 3. Experimental Design

To evaluate the capability of LLMs to analyze videos from a legal perspective, we collected a dataset of videos (section 3.1) and built an LLM pipeline to process these videos (section 3.2), in order to generate summaries of the videos (section 3.3) and draft legal letters (section 3.4), which are then evaluated quantitatively and qualitatively. The code to run these experiments is publicly available.[2]

### *3.1. Dataset*

Our dataset consists of 120 videos drawn from four legal domains. All videos were sourced from YouTube, as this platform provides publicly available material that captures how real people record and share footage of disputes and incidents in lived contexts, focused on everyday legal issues [8]. To ensure comparability across domains, we only included video with at least 2 seconds of footage where a factual legal problem can be seen, and that contained clear visual evidence to allow the factual evaluation.

Videos were selected at random from search results, provided they fit within one of the target domains (see Table 1). We also captured metadata for each video (see Table 2).

| Domain | Description | N Videos |
|---|---|---|
| Housing defects | Focused on everyday disputes and common defects in rental housing. Search terms included "apartment water leak,", "cockroach infestation apartment," "mold in rental apartment," "landlord refuses repairs," "broken heating." | 24 (20%) |
| Traffic incidents | Covered a range of vehicle accidents and road user conflicts. Search terms included "car crash compilation," "truck crash dashcam." | 42 (35%) |
| Property damage | Emphasized natural disasters and destruction. Search terms included "house fire aftermath," "{storm\|flood\|earthquake} damage to my house," "tenant destroyed property." | 34 (28.3%) |
| Product malfunctions | Targeted defective consumer goods and devices. Search terms included "batteries randomly catching on fire," "phone catches on fire," "batteries exploding." | 20 (16.7%) |
| **Total** | | **120 (100%)** |

**Table 1.** Domains, search strategies, and number of videos in the dataset.

### *3.2. Model*

To investigate the capabilities of AI to generate factual video summaries, we set up a pipeline that submits short video clips together with task-specific instructions. We used Gemini 2.5 Flash (stable version, last updated June 2025), which represents a state-of-the-art multimodal LLM suitable for processing video inputs. Google enables this model to understand videos by sampling images from the video at 1FPS, and audio at 1Khz. We likewise used functionality that allows the model to directly load videos from Youtube.[3]

We interacted with the model via its API, setting the temperature parameter to 0, and seed to 42. The specific prompts used are listed below. As recommended by the provider, we first passed the video to the model and then the textual prompt.

[2] Available at `https://github.com/maastrichtlawtech/jurix2025_LLM_video_analysis`

[3] `https://ai.google.dev/gemini-api/docs/video-understanding`

| Metadata Question | Result | N | % |
|---|---|---|---|
| Does the video contain a voice-over? | Yes | 69 | 57.5 |
| | No | 51 | 42.5 |
| What is the filming perspective of the video? | First Person | 48 | 40.0 |
| | Third Person | 37 | 30.8 |
| | Dash Cam | 35 | 29.2 |
| Does the video contain text in the footage? | Yes | 31 | 25.8 |
| | No | 89 | 74.2 |
| **What is the complexity level of the video?** | | | |
| **Level 1**: A single, clearly identifiable event occurs (e.g., a product breaking at a specific moment, water dripping from a ceiling). | | 58 | 48.3 |
| **Level 2**: Multiple events occur, but they remain individually distinguishable (e.g., a tenant showing several instances of mold damage in sequence). | | 60 | 50.0 |
| **Level 3**: Many events occur at once, making it difficult to distinguish the legally relevant facts (e.g., a chaotic traffic accident with several vehicles and bystanders). | | 2 | 1.7 |

**Table 2.** Summary of metadata characteristics across the video dataset.

### *3.3. E1 - Video Summaries*

In order to answer RQ1 and RQ2, we used Gemini to write short summaries of a video, to emulate a situation where a layperson has filmed a short clip and wants to use AI to help them fill out information in, e.g., a form. The prompt used was:

| |
|---|
| **Prompt:** I will give you a video showing some sort of legal issue. Write a 2-3 sentence summary of the legally relevant elements shown in the video, which will be used as part of a legal form or insurance claim. Make sure to only include facts that are part of the video. Be neutral and specific. |

**Table 3.** Prompt for Experiment 1 - Creation of video summaries

Once the summaries were generated, we assessed their quality through an annotation process. We drafted annotation guidelines and asked annotators (co-authors of this paper) to annotate each video with the criteria explained in Table 6. In total, seven annotators worked on the annotation, and each video was annotated by two. Annotators were instructed to ignore external knowledge, assumptions, or legally relevant inferences that were not explicitly visible in the video. In case of disagreements, a single annotator reconciled the annotations in taking into account the notes written by the initial annotators.

### *3.4. E2 - Legal letter drafts*

To answer RQ3, we also wanted to assess whether an LLM can flag legally relevant issues and link them to relevant legislation. This mimics a situation where a layperson asks an LLM to draft a full legal complaint letter based on a video. We selected 5 videos from each category and asked the model to draft a letter under dutch law. The prompt can be seen in Table 4. To ensure consistency and legal relevance, we specified who the addressee of the letter should be for each domain, see table 5. The resulting complaint letters were then qualitatively analyzed by legal experts to determine whether the model correctly identified and articulated core aspects of legal reasoning, such as identification of violated norms, damages and remedies claimed.

| |
|---|
| **Prompt:** Assume that the video happens in the Netherlands. Write a legal complaint letter from the perspective of the person filming the video, referring to relevant legal rules in the Netherlands. Write the letter in English. Address the letter to the {addressee}. |

**Table 4.** Prompt for Experiment 2 - Drafting of legal letters

| **Domain** | **Addressee** |
|---|---|
| Housing defects | Landlord |
| Property damage | Person who caused the damage |
| Traffic | Car that caused the crash |
| Product malfunction | Product producer |

**Table 5.** Mapping of domains to the relevant addressee.

## 4. Results

### *4.1. E1 - Video Summaries*

To answer RQ1 and RQ2, we generated summaries for 120 video clips and evaluated them based on overall quality, completeness and factuality. Table 6 shows the overall results of the rated video summaries. Overall, 37.5% of the summaries were rated as being of "high" quality, while 34.2% were ranked as "medium" and 28.3% were rated as being of low quality. 64.2% were found to be complete, while 55% were found to be factual. Figure 2 shows how the quality of the generated summaries varies by the various metadata (see section 3.1). Figure 3 shows how the completeness and factuality metrics vary based on the metadata.

| **Question / Criteria** | **N** | **%** |
|---|---|---|
| **Overall Quality of AI Summary** | | |
| **High**: All facts captured, no hallucinations. | 45 | 37.5 |
| **Medium**: Some facts missing or small additions, but without essential impact on the meaning of the footage. | 41 | 34.2 |
| **Low**: Key facts missing and/or hallucinations that misrepresent the event and drastically change its meaning. | 34 | 28.3 |
| **Completeness: Were all legally relevant occurrences visible in the video captured in the summary?** | | |
| **Yes**: The summary captured all legally relevant occurrences. | 77 | 64.2 |
| **No**: Some legally relevant occurrences were missing. | 43 | 35.8 |
| **Factuality: Were all of the facts in the description accurate to the video (hallucination-free)?** | | |
| **Yes**: The summary was factually accurate and hallucination-free. | 66 | 55.0 |
| **No**: The summary contained hallucinated or factually incorrect details. | 54 | 45.0 |

**Table 6.** Overall quality, completeness, and factuality of AI-generated summaries.

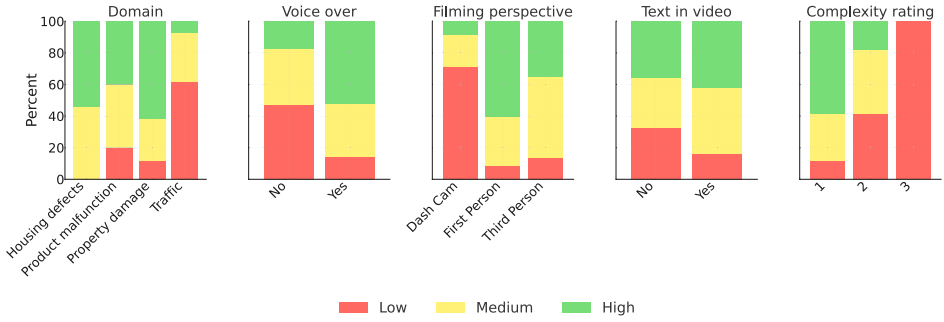



**Figure 2.** Overall quality of AI summaries across metadata factors (domain, voice-over, perspective, text, complexity).

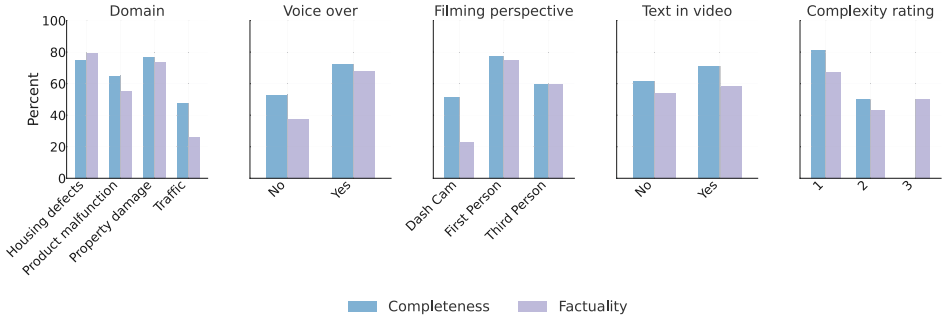



**Figure 3.** Completeness and factuality of AI summaries across metadata factors (domain, voice-over, perspective, text, complexity).

### *4.2. E2 - Legal Letter Drafts*

To answer RQ3, we used an LLM to draft legal letters based on 5 videos for each domain, (20 in total) and qualitatively evaluated the results. In general, the letters demonstrated structural clarity, references to relevant provisions, and a degree of legal reasoning that makes them functional for preliminary use. However, certain limitations emerged, particularly with respect to contextual accuracy and consistency in the application of legal standards. A more detailed analysis of these results will be presented in Section 5.3.

## 5. Discussion

Let us examine what the results imply about the research questions.

### *5.1. RQ1 - Can LLMs reliably summarize legally relevant facts from video footage?*

E1 indicates that the video summaries are promising, with 71.7% of the summaries being rated as of high or medium quality. Overall, 64.2% of the generated summaries captured all relevant occurrences, while 55% of the videos were found to be free of hallucinations.

Considering the difficulty of the task of observing a video with sound and summarizing it in a coherent manner that captures the relevant legal details, these results are

promising. However, they also highlight that the tool could not reliably generate accurate summaries. If deployed today, one would need a human in the loop to correct them. That said, being presented with a draft summary could reduce the time needed to draft a legal submission for lawyers, and present a useful first draft even for laypeople, who could correct factual mistakes but benefit from the tone and wording of the summaries.

We would also like to highlight the difficulty of annotating the videos. In the end, many of the questions saw significant disagreement between the annotators. Quality and completeness saw especially high disagreements. Partially, this may be resolved with more precise instructions of what should be covered in each summary or a different prompting style, which we will explore in future work. However, the question of which fact in a video is "legally relevant" is also quite ambiguous, and may lead to different answers between different legal experts as well. Once more, this research points to the fact that LLMs are at a level where evaluation against a gold-standard dataset is starting to become very difficult (see also [19,15,34,14]).

### *5.2. RQ2 - How do video characteristics influence the quality of LLM outputs?*

**Domain**: The domain had a large impact on the quality of the summaries (Figure 2):

*Housing defects:* Housing defects had the highest score out of all of the domains. Regarding completeness, the generated summaries were not specific enough in e.g. the size of mold or that water was wastewater. The hallucinations centered on being too confident in certain facts (i.e. bugs being cockroaches), spotting things not present in the video (e.g. debris or aerial spores) and over-relying on the narration.

*Product malfunction*: Overall quite reasonable scores. However, the model was bad at understanding sequences of events, especially when in rapid succession (see discussion regarding frame rate below). Here, the summary would make up things happening (such as objects being thrown or people being propelled by an explosion).

*Property damage*: Second-highest scoring category. The model sometimes over-relied on the voice-over, and sometimes ignored it. It likewise had issues in assessing the height of water level and would often skip over the general messiness of a room.

*Traffic*: By far the lowest scoring category. The model seems unable to capture sequences of events when the car was moving, and instead fully hallucinates what happens. This could stem from the fact that the model only receives a single frame per second. Future work will need to explore increasing the framerate. Even so, the model would often confuse directions, such as left and right, or passenger-side versus driver-side. Emergency vehicles and signs were also hallucinated in some instances.

**Voice-over**: Videos with a voice-over had higher quality ratings, completeness and factuality scores. The model was able to understand spoken voices in Dutch as well as English, and weaved the information into the description. In some instances, this caused hallucinations, since the model would believe what the voice said even if it was not backed up in the video, which may not be desirable in all circumstances.

**Perspective**: Dash cam, which was exclusively used in traffic videos, performed poorly. First person performed a bit better than third-person videos, perhaps because the person in front of the camera adds another element for the LLM to consider.

**Text**: Just like voice-over, text in videos did improve performance. The summaries often included date and times from overlays, and relied on the text.

**Complexity**: Videos of higher complexity were harder for the LLM to accurately summarize. Note that there are only 2 samples of complexity level 3.

### *5.3. RQ3 - To what extent can LLMs link these facts to the relevant legal rules?*

In E2, we asked the LLM to go further than merely summarizing the videos, by also drafting a legal complaint letter under Dutch law, which we then qualitatively evaluated. The legal rules differ by domain and letter addressee (see section 3.4).

#### *5.3.1. Housing Defects*

In the domain of housing defects, AI-generated letters often identified relevant statutory provisions and proposed potential remedies, providing a structured draft suitable for preliminary use. However, significant limitations were evident. The drafts consistently omitted reference to the Rental Defects Guide, a source of soft law that, although not formally binding, is systematically relied upon by courts and rental tribunals in such disputes. Furthermore, while the prescribed response period was often set at 14 days, in line with Dutch case law, some drafts specify significantly shorter deadlines, which would not be upheld in court. In addition, there was variability in the framing of remedies; for instance, the letters often did not specify amounts, which, given the nature of the video evidence, is understandable. Overall, while these drafts may serve as a starting point, they require substantial revision and careful expert oversight before being used in practice.

#### *5.3.2. Product Malfunction*

In cases involving product malfunctions, AI-generated letters generally relied on strict liability provisions, such as product liability statutes (e.g., Articles 6:185–6:186 DCC), and highlighted core principles such as consumer safety expectations. While the drafts suggested remedies and response deadlines that align with legal practice, a notable limitation was the consistent omission of complementary legal claims, particularly contractual non-conformity under Article 7:17 DCC or the general provision of tort (art. 6:162 DCC). This omission reduces the legal strategy of the letters, as both the producer and the seller could otherwise be held liable. In addition, the damage was not described in detail (no amounts are specified) resulting in letters that request acknowledgment of a compensation for all losses, regardless of scale.

#### *5.3.3. Property damage*

In property damage cases, the quality of AI-generated letters depended on the intended purpose. When tasked with describing the issue and drafting correspondence for the landlord, the letters demonstrated some functional utility: they referred to relevant legislative provisions and identified possible follow-up actions. However, the drafts did not specify the items that were damaged nor the value of the losses incurred, limiting the precision and practical usability of the letters. Despite these limitations, the outputs may still serve as a useful starting point for discussion and further refinement by legal professionals.

#### *5.3.4. Traffic*

In traffic-related cases, AI-generated letters again showed structural clarity and the ability to correctly identify and reference statutory provisions. However, factual inaccuracies were common, and because legal analysis relies on the stated facts, these errors directly undermined the correctness of the legal reasoning. This illustrates how deficiencies in factual accuracy "bleed through" to the legal assessment, emphasizing the necessity of

thorough human review. Consequently, the drafts are largely unusable. Additionally, the letters tended to focus on alleged traffic law violations, which may appear intuitively relevant but are legally inappropriate as a basis for civil claims between private individuals; whether a violation occurred is determined by criminal court.

#### *5.3.5. Conclusion*

Overall, the draft letters demonstrate structural clarity, the ability to reference relevant statutory provisions, and preliminary legal reasoning, making them potentially useful as starting points for drafting. However, usability is limited by recurring shortcomings, including factual inaccuracies, omissions of complementary legal claims, inconsistent or incomplete specification of damages, the omission of relevant soft law, and variability in proposed response periods. In particular, factual inaccuracies "bleed through" to the legal assessment, undermining the validity of the reasoning in mostly traffic disputes. This indicates that, while AI can assist in the drafting process and may be useful as a starting point for discussion, careful human review is still essential.

### *5.4. Toward a decision support app analyzing videos*

Finally, let us explore the question of whether and how the current state of video analysis using LLMs could be used in a legal decision support tool, as described in section 2.

Despite the unreliability of the generated summaries, we think they could serve as a useful step to drafting a description, especially in domains such as housing defects. The user should be informed to take a stable video that clearly shows the issue, and potentially describe what they see as well. Ideally each clip should show a single issue.

Even so, the text should be viewed and corrected by a legal professional, before being entered into a final form. The tool may likewise be used as an aid for identifying potential eligibility for compensation or legal recourse. In this capacity, it could help users recognize when they might have a valid claim and guide them toward appropriate next steps, such as contacting legal aid services or consumer protection organizations.

Regarding the drafting of legal letters, the models were able to draft legally sound content drawing from Dutch law, which is a quite small jurisdiction. However, the content was not perfect and also suffered from the inaccuracy in summarizing the videos. Integrating the video-recognition system into an existing legal information pipeline, such as the approach shown in [36], may address some of these shortcomings.

## 6. Conclusion & Future Work

We investigated the ability of LLMs to summarize videos from a legal perspective, and draft legal letters. The results are promising, but have room for improvement in terms of completeness and factuality, especially for complex videos with significant movement. We explored ways to mitigate the issues and opportunities for access to justice.

Future work will focus on enhancing the prompting (e.g. through chain-of-thought, reasoning models and RAG) and evaluation of the pipeline, experimenting with different video frame-rates, integration in neuro-symbolic systems and user interfaces, testing with users, and investigating more complex videos covering e.g. financial crimes or events from multiple perspectives.

## References


[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[2] Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. Can GPT-3 perform statutory reasoning? *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, page 22–31, 2023.

[3] Karl Branting, Carlos Balhana, Craig Pfeifer, John S Aberdeen, and Bradford Brown. Judges are from mars, pro se litigants are from venus: Predicting decisions from lay text. In *JURIX*, pages 215–218, 2020.

[4] Karl Branting and Sarah McLeod. Narrative-driven case elicitation. In *AI4AJ@ ICAIL*, 2023.

[5] L Karl Branting. Advisory systems for pro se litigants. In *Proceedings of the 8th international conference on Artificial intelligence and law*, pages 139–146, 2001.

[6] Ab Currie. The legal problems of everyday life. In *Access to justice*. Emerald Group Publishing Limited, 2009.

[7] Jakub Drápal, Hannes Westermann, and Jaromir Savelka. Using large language models to support thematic analysis in empirical legal studies. *arXiv preprint arXiv:2310.18729*, 2023.

[8] Trevor CW Farrow, Ab Currie, Nicole Aylwin, Lesley Jacobs, David Northrup, and Lisa Moore. Everyday legal problems and the cost of justice in canada: Overview report. *Osgoode Legal Studies Research Paper*, (57), 2016.

[9] Elisabetta Fersini, Gaia Arosio, Enza Messina, Francesco Archetti, and Daniele Toscani. Multimedia summarization in law courts: An environment for browsing and consulting. In *ICT4Justice*, pages 73–80, 2009.

[10] Nick Goodson and Rongfei Lu. Intention and context elicitation with large language models in the legal aid intake process, 2023.

[11] Morgan Gray, Jaromir Savelka, Wesley Oliver, and Kevin Ashley. Using llms to discover legal factors. *arXiv preprint arXiv:2410.07504*, 2024.

[12] Morgan A Gray, Jaromir Savelka, Wesley M Oliver, and Kevin D Ashley. Empirical legal analysis simplified: reducing complexity through automatic identification and evaluation of legally relevant factors. *Philosophical Transactions of the Royal Society A*, 382(2270):20230155, 2024.

[13] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14271–14280, 2024.

[14] Samyar Janatian, Hannes Westermann, Jinzhe Tan, Jaromir Savelka, and Karim Benyekhlef. From text to structure: Using large language models to support the development of legal expert systems. In *Legal Knowledge and Information Systems*, pages 167–176. IOS Press, 2023.

[15] Sayash Kapoor, Peter Henderson, and Arvind Narayanan. Promises and pitfalls of artificial intelligence for legal applications. *arXiv preprint arXiv:2402.01656*, 2024.

[16] Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254, 2024.

[17] Daniel Martin Katz et al. GPT-4 passes the bar exam. *SSRN 4389233*, 2023.

[18] Min Jung Lee, Dayoung Gong, and Minsu Cho. Video summarization with large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18981–18991, 2025.

[19] Megan Ma, Brandon Waldon, and Julian Nyarko. Conceptual questions in developing expert-annotated data. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 427–431, 2023.

[20] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

[21] Julie Macfarlane. The national self-represented litigants project: Identifying and meeting the needs of self-represented litigants final report. 2013.

[22] Ha-Thanh Nguyen, Randy Goebel, Francesca Toni, Kostas Stathis, and Ken Satoh. Black-box analysis: GPTs across time in legal textual entailment task. *arXiv preprint arXiv:2309.05501*, 2023.

[23] Amy Salyzyn, Lori Isaj, Brandon Piva, and Jacquelyn Burkell. Literacy requirements of court documents: An under-explored barrier to access to justice. *Windsor Yearbook of Access to Justice*, 33(2):263–301, 2016.

[24] Laura Savage and Susan McDonald. Experiences of serious problems or disputes in the canadian provinces, 2021. *Juristat: Canadian Centre for Justice Statistics*, pages 1–28, 2022.
[25] Jaromir Savelka, Kevin D Ashley, Morgan A Gray, Hannes Westermann, and Huihui Xu. Can GPT-4 support analysis of textual data in tasks requiring highly specialized domain expertise? *Proceedings of ASAIL'23*, 2023.
[26] Jaromir Savelka, Kevin D Ashley, Morgan A Gray, Hannes Westermann, and Huihui Xu. Explaining legal concepts with augmented large language models (GPT-4). *arXiv preprint arXiv:2306.09525*, 2023.
[27] Noel Semple. The cost of seeking civil justice in canada. *Can. B. Rev.*, 93:639, 2015.
[28] Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. Audio-visual llm for video understanding. *arXiv preprint arXiv:2312.06720*, 2023.
[29] Quinten Steenhuis, David Colarusso, and Bryce Willey. Weaving pathways for justice with gpt: Llm-driven automated drafting of interactive legal applications, 2023.
[30] Quinten Steenhuis and Hannes Westermann. Getting in the door: Streamlining intake in civil legal services with large language models, 2024.
[31] Quinten Steenhuis, Bryce Willey, and David Colarusso. Beyond readability with ratemypdf: A combined rule-based and machine learning approach to improving court forms. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 287–296, 2023.
[32] Richard Susskind. Online courts and the future of justice. 2019.
[33] Jinzhe Tan, Hannes Westermann, and Karim Benyekhlef. ChatGPT as an artificial lawyer? *Artificial Intelligence for Access to Justice (AI4AJ 2023)*, 2023.
[34] Jinzhe Tan, Hannes Westermann, Nikhil Reddy Pottanigari, Jaromír Šavelka, Sébastien Meeùs, Mia Godet, and Karim Benyekhlef. Robots in the middle: Evaluating llms in dispute resolution, 2024.
[35] Hannes Westermann. *Using artificial intelligence to increase access to justice*. PhD thesis, Université de Montréal, October 2023.
[36] Hannes Westermann. Dallma: Semi-structured legal reasoning and drafting with large language models. *2nd Workshop on Generative AI and Law*, 2024.
[37] Hannes Westermann and Karim Benyekhlef. JusticeBot: A Methodology for Building Augmented Intelligence Tools for Laypeople to Increase Access to Justice. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ICAIL '23, pages 351–360, New York, NY, USA, September 2023. Association for Computing Machinery.
[38] Hannes Westermann and Karim Benyekhlef. Justicebot: A methodology for building augmented intelligence tools for laypeople to increase access to justice. *ICAIL 2023*, page 351–360, 2023.
[39] Hannes Westermann et al. Bridging the gap: Mapping layperson narratives to legal issues with language models. In *ASAIL 2023*, 2023.
[40] Hannes Westermann and Jaromir Savelka. Analyzing images of legal documents: Toward multi-modal llms for access to justice, 2024.
[41] Hannes Westermann, Jaromir Savelka, and Karim Benyekhlef. LLMediator: Gpt-4 assisted online dispute resolution. *Artificial Intelligence for Access to Justice (AI4AJ 2023)*, 2023.
[42] Fangyi Yu, Lee Quartey, and Frank Schilder. Legal prompting: Teaching a language model to think like a lawyer, 2022.
[43] John Zeleznikow. Using web-based legal decision support systems to improve access to justice. *Information & Communications Technology Law*, 11(1):15–33, 2002.
[44] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251581*

# Which Neurons Nudge Normative Stance? Causal Tests and Mechanistic Evidence via Contrastive Last-Token Steering

Davide LIGA [a,1], Liuwen YU [a]
[a] *University of Luxembourg*
ORCiD ID: Davide Liga https://orcid.org/0000-0003-1124-0299, Liuwen Yu https://orcid.org/0000-0002-7200-6001

**Abstract.** Normative stance underlies decisions in law, legal reasoning, policy, and safety-critical settings. A model's judgment of what is permissible vs. impermissible often determines its downstream behavior. We study how to steer a language model's normative stances at inference time by adding a tiny, contrastive perturbation to the *last-token* neural activation in late MLP layers (contrastive last-token steering). For each normative prompt, we construct a contrast direction by comparing its last-token activation to that of a minimally edited variant that implies a more permissive normative stance (e.g., "acceptable" rather than "wrong"). During generation, we add this vector at the last token; a single strength parameter $\alpha$ controls how strongly and in which direction we push the model's stance (permissive vs. restrictive). Impact is measured as the change in a next-token logit margin between permissive and restrictive continuations. To avoid overclaiming, we calibrate a threshold $\tau$ on neutral controls (same layers, tempered strengths with $|\alpha| \leq 1$) and count success only when the shift exceeds $\tau$ in the expected direction. We also assess *specificity* by verifying that, on neutral control prompts, steered outputs exactly match unsteered baselines. Beyond component-level tests, we probe *neuron-level locality* by steering only the top-$k$ contrastive neurons (ranked by last-token contrast) and confirming reversibility on our test set: $+\alpha$ produces the shift and $-\alpha$ reverses it. The method is training-free, uses standard forward hooks, and we report pilot results on Llama-3-8B-Instruct.

**Keywords.** Large Language Models, Normative Alignment, Normative Reasoning, Mechanistic Interpretability

## 1. Introduction

Large language models (LLMs) now routinely engage in tasks that require moral and normative judgment—choosing between conflicting values, resolving ethical and *legal* dilemmas, and aligning with user-defined principles. LLMs are increasingly assessed and even used around morally salient judgments. When prompted with trolley-style scenar-

[1]Corresponding Author: Davide Liga, davide.liga@uni.lu.

ios or everyday social dilemmas, these models can articulate moral reasoning and produce graded or binary decisions [1,2]. At the same time, moral judgments are highly sensitive to prompt framing [3] and exhibit cultural biases and limits [4,5]. Some experiments show that ChatGPT's moral advice can be inconsistent yet still shifts users' decisions [6], and LLMs can outperform humans on social situational judgment tests [7]; moreover, people may rate LLMs' ethical advice as rivaling a professional ethicist [8]. Yet while LLMs' outputs can be coherent, the internal mechanisms that steer these normative choices remain opaque. Understanding which neural components encode normative preferences allows us to explore the alignment of LLMs with norms and to ensure that automated systems remain accountable and transparent in their normative reasoning (which is crucial in the legal domain). In this paper, we ask the following research question: Which internal representations are responsible for a model's moral stance, and can we *causally* intervene on them without retraining?

**Terminology.** We call a stance *permissive* when continuations such as *"acceptable"/"justified"* are preferred over *"wrong"/"unacceptable"*, and *restrictive* otherwise. *Specificity* means the intervention leaves unrelated outputs unchanged; we measure it as canonicalized exact–match of steered vs. unsteered greedy generations on neutral controls. For example, the minimal pair *"Lying to save lives is morally ..."* vs. *"Lying to save face is morally ..."* implies opposite stances under small textual changes (*mutatis mutandis*, this is analogous to minimal, but substantial fact variations in case law that may flip a legal judgment).

**Our approach.** We study these questions via *contrastive activation steering*. Rather than swapping whole activations, we compute a *direction of change* between a minimally different prompt pair that elicits opposite stances ("permissive" vs. "restrictive"). We then inject a small, controlled perturbation along this contrast direction at the *last input position* inside selected MLP blocks (primarily late layers) during the forward pass. The intervention is scaled by a single real parameter $\alpha$ (the *strength*); positive $+\alpha$ is aligned—by a brief sign probe—to push toward the stance predicted by the counterfactual prompt, while negative $-\alpha$ reverses the effect. We quantify causal impact using a calibrated *next-token logit margin* between mutually exclusive continuations corresponding to the two stances. For moral pairs we always steer the last token; for neutral controls we keep the same last-token position but use masking to check that outputs remain unchanged.

**Evaluation frame.** To avoid over-interpreting single runs, we adopt a compact but principled evaluation pipeline:

1. **Directional validity.** At a given $\alpha$, does the decision shift $\Delta$ have the expected sign (toward permissive for $+\alpha$, toward restrictive for $-\alpha$) as predicted by the counterfactual pair?
2. **Calibrated decision threshold.** We estimate a threshold $\tau$ from neutral control prompts using the same layers and a tempered subset of strengths ($|\alpha| \leq 1$, i.e., $\alpha \in \{-1.0, -0.8, 0.8, 1.0\}$), and count a *success* only when $\Delta \cdot \text{sign expected}$ exceeds $\tau$.
3. **Specificity.** We test that steering leaves unrelated outputs unchanged by comparing steered vs. unsteered generations on neutral controls (canonicalized exact match), reporting preservation and corruption rates.
4. **Dose–response and reversibility.** We sweep $\alpha$ over a symmetric grid (positive/negative). Neuron-level tests include explicit reversibility checks ($+\alpha$ vs.

$-\alpha$). At the component level we report success across strengths; a formal monotonicity score is left for future work.

5. **Neuron-level locality.** Within late MLP layers, we select top-$k$ neurons by last-token contrast magnitude and test whether small, sign-aligned subsets can reproduce (and reverse) the steering effect.

**Contributions.** We present a compact pipeline for causally steering moral stances in pretrained LLMs, showing mechanistic evidence that paves the way for an integration between the field of AI&Law and the young field of *Mechanistic Interpretability*:

1. **Contrastive last-token steering.** A causal intervention that targets *contrastive, last-token* directions by modifying MLP outputs that feed the residual stream (we emphasize late layers but also test selected earlier layers). The method requires no fine-tuning and uses standard forward hooks during inference only.
2. **Calibrated logit-margin metric.** A decision-shift metric based on the change in logit margin between opposed completions, with a neutral-control–derived threshold for consistent comparisons across prompts, layers, and steering strengths.
3. **Neuron-level locality.** A neuron-level analysis in late MLPs showing that small, sign-aligned subsets (top-$k$ by contrast) can flip stance on our test set and exhibit reversibility with $+\alpha/-\alpha$.

Together, these components provide a reproducible workflow for measuring and controlling moral decisions in LLMs using only inference-time interventions.

**Scope.** We evaluate on Meta-Llama-3-8B-Instruct and a set of moral minimal pairs; claims are about steering effects within this setting. Component-level dose–response is reported via strength sweeps; explicit monotonicity scoring is left as future work.

**Paper organization.** Section 2 contextualizes this work in the literature. Section 3 presents the contrastive last-token steering method. Section 4 details the experimental setup and reports results. Section 5 discusses limitations and future directions, and concludes this work.

## 2. Related Work

**Moral behavior in language models.** LLMs exhibit moral and social judgments that appear to reflect implicit values learned during pretraining. Prior work has evaluated these behaviors using curated benchmarks and ethical dilemmas. For example, [9] introduced Delphi to make ethical judgments across diverse scenarios, and [10] proposed ETHICS to assess dimensions such as justice, virtue, and utilitarianism. [11] analyze moral foundations in LLMs, while [12] explore self-refinement strategies that can influence responses to moral questions. Much of this literature is output-centered; the internal mechanisms producing these stances remain comparatively less characterized, despite their importance for domains such as computational legal reasoning, where explanations of normative stance are critical.

**Mechanistic interpretability and causal interventions.** Mechanistic interpretability aims to reverse-engineer computations inside transformers [13], including work on superposition and feature sharing [14] and on targeted weight edits such as ROME [15]. A complementary line of research uses *activation patching*/causal tracing to test how

swapping or modifying hidden states affects behavior. Our approach follows this causal-intervention paradigm but in a localized setting: we construct a *contrastive direction* from minimally different prompt pairs and add a small, inference-time perturbation at the *last token* inside late MLP blocks, leaving model weights unchanged.

**Probing vs. causal control.** Linear probes and concept vectors (e.g., CAVs; [16]) indicate that a representation correlates with a concept, but correlation does not by itself imply that manipulating that representation *causes* behavior to change. Bias and social-attribute evaluations (e.g., [17]) similarly diagnose tendencies without isolating mechanism. By contrast, we emphasize *causal* tests: we measure a calibrated change in a next-token logit margin when we add a small, sign-aligned perturbation to late-layer MLP activations. Beyond component-level tests, we examine *neuron-level locality* by steering only the top-$k$ contrastive neurons and verifying reversibility with $+\alpha/-\alpha$.

**Steering without fine-tuning.** Inference-time steering spans prompt-based control, activation editing, and representation arithmetic. Our contribution fits within this space but differs in three respects: (i) we derive *contrastive, last-token* directions from minimally changed moral pairs (rather than global directions or weight edits), (ii) we *calibrate* a decision threshold on neutral controls to avoid overclaiming spurious flips, and (iii) we validate *locality and reversibility* by showing that small neuron subsets in late MLPs can reproduce and reverse the effect. This yields an efficient, reproducible pipeline for causal moral steering using standard forward hooks at inference time.

**AI&Law.** Integrating symbolic and sub-symbolic AI is becoming increasingly important in the field of AI&Law [18], and Mechanistic Interpretability offers a fundamental opportunity to explore new integrations between these two traditional paradigms in AI. Our work is a first attempt to propose an integration between these two communities.

## 3. Method

We describe an interventional, inference-time procedure to steer a pretrained language model's moral stance by adding small, targeted perturbations to late MLP activations. The procedure has four components: (i) a *decision signal* defined as a next-token logit margin between mutually exclusive moral continuations; (ii) *contrastive last-token steering* that injects a direction computed from minimally different prompt pairs; (iii) *calibration and specificity* using neutral controls; and (iv) a *neuron-level* variant that targets small subsets of units.

**At a glance.** For each moral pair we (a) build a last-token direction from the base vs. source contrast (or a PCA fallback—last-token, position-aware—if unavailable), (b) orient it via a one-shot *sign probe* and inject it into selected late MLP layers at strength $\alpha$, and (c) score the change in the next-token margin against a calibrated threshold; we also (d) check specificity on neutral controls and (e) run a neuron-level variant targeting top-$k$ units. We focus on last-token interventions because late-layer MLP features tend to consolidate decision signals; in practice our component-level run steers a fixed set of MLP layers spanning early through late blocks, while neuron-level tests focus on late layers where the effect is most localized. The steering strength $\alpha$ is not learned: we sweep a symmetric grid and select $\alpha$ by a calibrated success metric, with the threshold $\tau$ estimated from neutral controls using the same layers and tempered strengths.

### 3.1. Task and Decision Signal

We work with *minimal pairs* of prompts $(P_{\text{base}}, P_{\text{src}})$ that differ by a small surface change but imply opposite stances ("permissive" vs. "restrictive"). Let $x$ be the tokenized input for $P_{\text{base}}$, and let $y_+$ and $y_-$ denote mutually exclusive next-token continuations aligned with "permissive" and "restrictive" readings, respectively.[2] Given next-token logits $\ell(\cdot)$, we define the *logit margin* as $m = \ell(y_+) - \ell(y_-)$, and the *decision shift* for an intervention as $\Delta = m_{\text{patched}} - m_{\text{base}}$. Positive $\Delta$ indicates movement toward a permissive decision.

### 3.2. Contrastive Last-Token Steering

Let $h^\ell(x) \in \mathbb{R}^{T\times H}$ be the MLP output at layer $\ell$ for input $x$ (sequence length $T$, hidden size $H$). We steer only the *last token*. Define the last-token shorthand $h_T^\ell(P) \equiv h^\ell(P)_{T,:}$, and form a per-pair *contrastive direction* by differencing the base and counterfactual prompts:

$$\Delta h_T^\ell := h_T^\ell(P_{\text{src}}) - h_T^\ell(P_{\text{base}}),$$

$$v_\ell := \frac{\Delta h_T^\ell}{\|\Delta h_T^\ell\|_2 + \varepsilon},$$

where $\varepsilon$ is a small device/dtype-safe constant.

At inference time, inside a selected set of MLP layers $\mathscr{L}$, we add a tiny perturbation at the last position: $\tilde{h}_T^\ell = h_T^\ell + \alpha\, s_\ell\, \sigma_T^\ell\, v_\ell, \qquad \ell \in \mathscr{L}$,where $\alpha \in \mathbb{R}$ is the *steering strength*, $\sigma_T^\ell := \text{std}(h_T^\ell)$ is the local hidden-state scale (std. across the hidden dimension), and $s_\ell \in \{+1, -1\}$ orients the effect so that $+\alpha$ increases the permissive margin implied by $P_{\text{src}}$. We set $s_\ell$ with a one-shot sign probe: apply a small $+\alpha_{\text{probe}}$ (e.g., 0.8) on that layer and pair; if the measured margin shift $\Delta$ is opposite to the expected direction, flip the sign. When steering multiple layers simultaneously, we scale by $1/\sqrt{|\mathscr{L}|}$ to keep the intervention magnitude comparable.

**Fallback direction.** We compute a small, position-aware PCA "moral subspace" by applying PCA to last-token MLP activations across moral prompts and ranking components by how well their projections predict a permissive–restrictive margin. When a per-pair contrast is unavailable, we use the normalized average of the top-ranked PCA directions as a fallback and orient it with the same sign probe.

### 3.3. Calibration and Specificity

We estimate a *decision threshold* $\tau$ from neutral control prompts (e.g., weather/geography). We apply the same layers and a tempered subset of strengths ($|\alpha| \leq 1$) to controls and set $\tau$ to the empirical 95$^{\text{th}}$ percentile of $|\Delta|$ aggregated over those controls and strengths. A trial on a moral pair is counted as a success only if

$$\Delta \cdot \text{sign}_{\text{expected}} > \tau,$$

[2] In practice, we pool a small set of stance tokens/phrases such as *right/acceptable/justified* vs. *wrong/unacceptable/unjustified*.

where $\text{sign}_{\text{expected}} \in \{+1, -1\}$ encodes the stance implied by the counterfactual source, computed by either a small NLI-based scorer or a lexical heuristic over the source completion.[3] All moral-pair interventions target the last token (`gate_on_mask=False`). For neutral controls we also gate the hook at the last token (`gate_on_mask=True`); specificity holds when steered control outputs exactly match their unsteered baselines. **Specificity metric:** we test that steering leaves unrelated content unchanged by re-generating short outputs for neutral controls with and without hooks and reporting *control preservation* (canonicalized exact match) and *corruption* rates.

### 3.4. Neuron-Level Steering

For each $\ell \in \mathscr{L}$, we select top-$k$ neurons by absolute contrast magnitude at the decision position: $S_k^\ell = \text{top-}k\big(|h^\ell(P_{\text{src}})_{T,:} - h^\ell(P_{\text{base}})_{T,:}|\big)$.We then add per-neuron nudges at the last token: $\tilde{h}_{T,j}^\ell = h_{T,j}^\ell + \alpha \cdot s_j^\ell \cdot \sigma_T^\ell, \qquad j \in S_k^\ell$,with signs $s_j^\ell$ aligned by a single-neuron probe (flip if $+\alpha_{\text{probe}}$ yields a negative $\Delta$). We evaluate $k \in \{1, 5, 10\}$ and test *reversibility* under $-\alpha$. The neuron-level success threshold is $\tau_{\text{neuron}} = \max(0.5\tau, 0.02)$ to reflect smaller expected effect sizes.

### 3.5. Implementation Notes

All interventions are implemented with standard hooks (no parameter updates): *forward hooks* for component-level steering and *forward pre-hooks* for neuron-level nudges. We target MLP modules of a decoder-only transformer, steer only the last token, normalize steering vectors with a device/dtype-safe $\varepsilon$ (allocated on the module's device and dtype), and scale by the local std. to reduce sensitivity to layer/dtype. We use greedy decoding for short baselines and read next-token logits directly to compute margins; the sweep selects the $\alpha$ with the highest calibrated success rate. An optional `split_variants` mode repeats the pipeline over disjoint prompt buckets for robustness.

## 4. Experimental Setup and Results

We now specify the concrete settings used to run the method in Section 3. This section details experimental setup and results. We use **Meta-Llama-3-8B-Instruct** as the target model (decoder-only transformer). Unless otherwise noted, we run in `torch.float32` with greedy decoding (`do_sample=False`) and a fixed random seed (`seed=42`). Experiments were executed on two environments: a local Apple Silicon macOS device with Metal Performance Shaders (MPS) backend, and a Colab (A100) environment with CUDA and with bfloat16/float16 mixed precision for the base model (steering math (norms, scaling) uses the module's compute dtype with device-allocated $\varepsilon$ to remain dtype-safe). We register forward hooks only; model weights are never updated.

[3] *Implementation note.* The NLI scorer is only used to set the expected sign for orientation; the primary outcome is the next-token logit-margin shift.

### 4.1. Minimal-Pair Prompts

We evaluate on a set of *moral minimal pairs* that differ by a short phrase but imply opposite stances (permissive vs. restrictive). Examples include:

- *"Lying to save lives is morally ..."* vs. *"Lying to save face is morally ..."*
- *"Helping others is morally ..."* vs. *"Harming others is morally ..."*
- *"Stealing to save lives is ..."* vs. *"Stealing for personal gain is ..."*
- *"Breaking rules to help others is ..."* vs. *"Breaking rules to help yourself is ..."*

We instantiate $n=284$ minimal pairs by combining seed contrasts with a deterministic paraphrase (e.g., *morally→ethically*, insert "is generally") and two light context wrappers (*In general,* / *From an ethical standpoint,*), then deduplicating and shuffling.[4] For each pair $(P_{\text{base}}, P_{\text{src}})$, we generate a short greedy completion for $P_{\text{src}}$ and use it to set the expected direction (Section 3.3) via a small NLI scorer or a lexical heuristic.

### 4.2. Control Prompts

To calibrate the decision threshold and measure specificity, we use neutral prompts unrelated to morality (e.g.: "The weather today is ..."; "Mathematics involves numbers and ..."; "The capital of France is ..."). For these prompts, we run the same hook configuration and compare steered vs. unsteered generations after canonicalization (lowercasing, symbol stripping). Control preservation is reported as exact-match rate.

### 4.3. Layers and Steering Vectors

We first run a lightweight diagnostic to identify influential components (position-aware PCA). For the component-level run we steer a fixed set of MLP layers spanning early and late blocks, while neuron-level tests focus on late layers. For the reported runs we steer *MLP* blocks at the following layers: $\mathscr{L} = \{0,4,8,10,16,18,20,22,28,31\}$(0-indexed). For each pair and layer $\ell \in \mathscr{L}$ we compute a last-token contrast vector $v_\ell$ (Section 3.2); when unavailable, we fall back to a position-aware PCA direction computed from last-token MLP activations. Vector orientation is aligned per layer by a one-shot sign probe ($\alpha_{\text{probe}} \approx 0.8$) so that $+\alpha$ increases the permissive–restrictive margin, and we scale by $1/\sqrt{|\mathscr{L}|}$ when steering multiple layers. Neuron-level tests focus on late layers (e.g., $\{28,31\}$). For moral pairs we always steer the last token; for neutral controls we use the same last-token position with mask gating so hooks are inert unless that position is active.

### 4.4. Strength Sweep and Decoding

We sweep a symmetric coarse grid $\alpha \in \{-1.2,-1.0,-0.8,0.8,1.0,1.2\}$,and scale the per-layer intervention by $1/\sqrt{|\mathscr{L}|}$ when steering multiple layers. Directions are oriented by a one-shot sign probe (see §3.2; $\alpha_{\text{probe}} \approx 0.8$) so that $+\alpha$ increases permissiveness. For each $\alpha$ and each pair, we run a single greedy step (`do_sample=False`) to read off next-token logits and compute the decision shift $\Delta$. The calibrated success across strengths is plotted in Figure 1. We also log the sign of the margin change (permissive vs. restrictive) for summary counts.

[4] Expansion is deterministic and we set `seed=42` before sampling and shuffling.

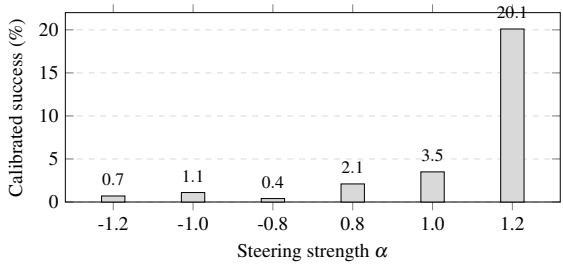



**Figure 1. Calibrated success across strengths.** Results on $n$=284 moral pairs, steering MLP layers $[0,4,8,10,16,18,20,22,28,31]$; threshold $\tau$=0.078 from neutral controls. "Success" is the fraction of pairs with $\Delta \cdot \text{sign}_{\text{expected}} > \tau$. Counts per $\alpha \in \{-1.2,-1.0,-0.8,0.8,1.0,1.2\}$: [2, 3, 1, 6, 10, 57].

### *4.5. Calibration Protocol*

We estimate the decision threshold $\tau$ using the neutral control prompts from Section 4.2, following the procedure in Section 3.3. Calibration uses the same layers $\mathscr{L}$ and a tempered subset of strengths ($|\alpha| \leq 1$). We compute $|\Delta|$ for every (control prompt, strength) combination and set $\tau$ to the empirical 95th percentile of this pooled set; this $\tau$ is then fixed for the run and used to decide success on moral pairs. For neuron-level tests we use $\tau_{\text{neuron}} = \max(0.5\,\tau, 0.02)$.

### *4.6. Neuron-Level Configuration*

For each $\ell \in \mathscr{L}$, we select top-$k$ neurons by absolute contrast at the last token with $k \in \{1,5,10\}$. We apply per-neuron nudges using forward *pre*-hooks at the last token, aligning each neuron's sign with a small probe. For $k = 1$ we additionally test *reversibility* by applying $-\alpha$ at the same magnitude that yielded a positive flip.

### *4.7. Evaluation Metrics*

We report the following:

- **Calibrated success rate:** percentage of pairs with $\Delta \cdot \text{sign}_{\text{expected}} > \tau$ at a given $\alpha$. We select the best $\alpha$ on the sweep.
- **Direction counts:** number of permissive vs. restrictive shifts (sign of margin change) at the best $\alpha$.
- **Control preservation / corruption:** exact-match rate on neutral controls (canonicalized) with/without hooks.
- **Neuron-level flips and reversals:** fraction of pairs that flip for $+\alpha$ (and flip back for $-\alpha$) at $k \in \{1,5,10\}$.

All metrics operate on next-token logits (no post-hoc classification of long completions).

### *4.8. Results*

We evaluate on $n$=284 moral minimal pairs and a small pool of neutral controls, steering MLP layers $[0,4,8,10,16,18,20,22,28,31]$ with a symmetric strength sweep. We report (i) a calibrated success rate, which counts a pair only if the next-token permis-

| Model | Layers | $\alpha^\star$ | $\tau$ | Success | Base → perm. | Steer → perm. | Steer → restr. | Spec. |
|---|---|---|---|---|---|---|---|---|
| Llama-3-8B -Instruct | [0,4,8,10,16, 18,20,22,28, 31] | 1.2 | 0.078 | 20.1% (≈57/284) | 253/284 | 278/284 | 6/284 | 100.0% (852/852) |

**Table 1. Steering summary at best strength.** $n$=284 moral pairs; three neutral controls per pair. $\tau$ is the 95th percentile of $|\Delta|$ measured on controls. Specificity (Spec.) is exact-match control preservation (steered control outputs equal unsteered baselines).

| Top-$k$ | Flips @ $+\alpha$ | Reversal @ $-\alpha$ | $n$ pairs |
|---|---|---|---|
| 1 | 89.1% | 88.7% | 284 |
| 5 | 100.0% | 100.0% | 284 |
| 10 | 100.0% | 100.0% | 284 |

**Table 2. Neuron-level locality in late MLPs (last token).** Fraction (and counts) of moral pairs ($n$=284) that flip under $+\alpha$ and reverse under $-\alpha$ when steering only the top-$k$ contrastive neurons in late layers $\{28, 31\}$. Evaluated at $\alpha^\star$=1.2 with $\tau_{\text{neuron}} = \max(0.5\,\tau, 0.02)$ and $\tau$=0.078 from neutral controls.

sive–restrictive margin shift exceeds the control-derived threshold $\tau$ in the expected direction, and (ii) the directional effect (how often the shift is permissive vs. restrictive regardless of magnitude). We also measure specificity on controls and test neuron-level locality. General results for our targeted LLM (Llama-3-8B-Instruct) are summarized in Table 1. **Component-level steering.** At the best strength on the sweep ($\alpha^\star = 1.2$), the calibrated success rate is **20.1%** (≈ **57/284**) under a control-derived threshold of $\tau = 0.078$. Directionally, **278/284** pairs shift permissive and only **6/284** restrictive, raising the permissive count from **253**→**278** (+8.8 percentage points). Specificity on neutral controls is **100%**: **852/852** exact matches and **0%** corruptions (three controls per moral pair; $3 \times 284 = 852$). Together these results show a reliable push in the intended direction with no detectable spillover to unrelated text. With $k$=5 and $k$=10, we observe flips on *all* pairs in this run and clean reversibility: applying $+\alpha$ produces the shift and $-\alpha$ cancels it. For $k$=1, effects are smaller and more variable (flips **89.1%**, reversals **88.7%**). See Table 2. These results indicate that a compact set of late-MLP units is sufficient to control the decision margin while preserving specificity. **Interpretation.** Thresholded success is conservative by design—many pairs move in the correct direction but remain below $\tau$. Neuron-level edits concentrate causal mass and avoid cross-layer cancellation, explaining their much higher flip and reversal rates compared to component-level mixing.

## 5. Conclusions and Limitations

***Limitations*** Our claims are bounded by several design choices. This section summarizes the most important limitations of our experimental design. **Prompt scope and labels.** We rely on *minimal pairs* that flip a permissive/restrictive stance by small textual edits. This operationalization is convenient for controlled experiments but does not cover the breadth of open-ended moral reasoning. Moreover, the polarity mapping ("permissive" vs. "restrictive") is induced via a small set of lexical or NLI templates; alternative label spaces (e.g., deontic vs. consequentialist justifications) are not tested. **Last-token locality.** All interventions are applied at the final input position. At the component level we steer a fixed set of MLP layers (including early/mid and late layers), while the neuron-level analysis focuses on late layers. This tests a specific hypothesis—that decisive evidence is consolidated at the decision token—but ignores earlier positions and

cross-token computations. Effects that depend on multi-sentence context may be underrepresented. **Metric narrowness.** We quantify shifts with a next-token *logit margin* between opposed continuations and use greedy decoding only for short baselines and control texts. This metric is simple and comparable across prompts but omits downstream decoding dynamics and multi-token rationales. The optional NLI scoring is template-based and model-dependent; it serves as a weak semantic check rather than a comprehensive evaluator. **Calibration sensitivity.** The decision threshold is estimated from a small set of neutral controls. The resulting $95^{\text{th}}$–percentile cutoff can vary with the control pool, strength grid, and decoding policy (we use greedy decoding during calibration). Broader, category-balanced controls would yield a more stable baseline. **Selection/tuning bias.** We select the best steering strength $\alpha$ by sweeping on the same set of moral pairs we later summarize. This can inflate apparent success. A stricter protocol would separate a tuning split from a held-out evaluation split or use nested cross-validation. **Evaluation scale.** Our evaluations use a moderate prompt set ($n$=284 moral minimal pairs) and a small pool of neutral controls. The pipeline emphasizes internal validity (directionality, reversibility, specificity), not exhaustive benchmarking. Broader claims would require larger and more diverse prompt pools, stronger statistics, and preregistered analysis plans. Specificity is measured as canonicalized exact-match of control generations (lowercasing and symbol stripping) under greedy decoding. **Model scope.** We evaluate a single model (Meta–Llama–3–8B–Instruct), so all claims are specific to this setting. Generalization to other models and variants remains open. In follow-up work we plan to vary (i) *size* (smaller/larger Llama-3 variants), (ii) *instruction-tuned vs. base* checkpoints, and (iii) *model family/architecture* (e.g., Llama vs. Qwen/Mistral; dense vs. MoE) to test whether the observed late-layer locality, directional effects, and specificity persist.

***Conclusions.*** We introduced a training–free, *contrastive last–token* steering method that modifies selected MLP outputs using directions derived from minimally different prompt pairs (evaluated here on Meta–Llama–3–8B–Instruct). At the component level (layers $[0,4,8,10,16,18,20,22,28,31]$), the calibrated success rate at the best strength ($\alpha^{\star}$=1.2) is 20.1% with $\tau$=0.078, while the directional effect is strong ($278/284$ permissive shifts) and control preservation is perfect ($852/852$). At the neuron level, steering top–$k$ late–layer units yields near–universal flips and clean reversals under $\pm\alpha$ (100% for $k$=5 and $k$=10). A calibrated logit–margin metric—with thresholds estimated from neutral controls—enables comparable measurement of directionality and effect size across prompts and strengths. In pilot neuron–level tests, small sign–aligned subsets in late layers reproduced and reversed stance shifts, while neutral controls remained largely unchanged. These observations indicate that, for our prompt set and model, parts of moral–stance formation can be *influenced* by small, interpretable perturbations at inference time, a finding with potential implications for legal AI systems that must balance transparency with normative alignment. By identifying which internal components influence normative outcomes, our approach advances transparency and accountability in normative AI systems, addressing concerns raised by the EU AI Act regarding opacity and bias in automated decision-making. Immediate next steps include: (i) scaling evaluations with larger and more diverse prompt pools, (ii) generalization tests via split variants and out–of–distribution pairs, and (iii) ablations and placebos (random/shuffled directions, layer swaps, and component zeroing) to strengthen causal claims. Longer term, span–wise steering and cross–model replication can probe how localized these decision circuits are and how they interact with decoding policies and rationale generation.

## Acknowledgement

Liuwen Yu thanks the FNR for their support through the SERAFIN project (C24/19003061 /SERAFIN) and the University of Luxembourg for the Marie Speyer Excellence Grant for the project Formal Analysis of Discretionary Reasoning (MSE-DISCREASON). Davide Liga was funded by the Fonds National de la Recherche (FNR), Luxembourg, under the D4H project (grant number PRIDE21/16758026).

## References

[1] Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, et al. The Moral Machine experiment. Nature. 2018;563(7729):59-64. Available from: https://www.nature.com/articles/s41586-018-0637-6.

[2] Zaim bin Ahmad MS, Takemoto K. Large-scale moral machine experiment on large language models. PLOS ONE. 2025;20(5):e0322776. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0322776.

[3] Oh S, Demberg V. Robustness of large language models in moral judgements. Royal Society Open Science. 2025;12(4):241229. Available from: https://royalsocietypublishing.org/doi/10.1098/rsos.241229.

[4] Touileb S, Nozza D. Measuring harmful representations in Scandinavian language models. arXiv preprint arXiv:221111678. 2022.

[5] Jiang L, Hwang JD, Bhagavatula C, Le Bras R, Liang JT, Levine S, et al. Investigating machine moral judgement through the Delphi experiment. Nature Machine Intelligence. 2025;7:145-60. Available from: https://www.nature.com/articles/s42256-024-00969-6.

[6] Krügel S, Ostermaier A, Uhl M. ChatGPT's inconsistent moral advice influences users' judgment. Scientific Reports. 2023;13(1):4569. Available from: https://www.nature.com/articles/s41598-023-31341-0.

[7] Mittelstädt JM, Maier J, Goerke P, Zinn F, Hermes M. Large language models can outperform humans in social situational judgments. Scientific Reports. 2024;14(1):27449. Available from: https://www.nature.com/articles/s41598-024-79048-0.

[8] Dillion D, Mondal D, Tandon N, Gray K. AI language model rivals expert ethicist in perceived moral expertise. Scientific Reports. 2025;15(1):4084. Available from: https://www.nature.com/articles/s41598-025-86510-0.

[9] Jiang L, Hwang JD, Bhagavatula C, Bras RL, Liang J, Dodge J, et al. Can machines learn morality? the delphi experiment. arXiv preprint arXiv:211007574. 2021.

[10] Hendrycks D, Burns C, Basart S, Critch A, Li J, Song D, et al. Aligning ai with shared human values. arXiv preprint arXiv:200802275. 2020.

[11] Abdulhai M, Serapio-Garcia G, Crepy C, Valter D, Canny J, Jaques N. Moral foundations of large language models. arXiv preprint arXiv:231015337. 2023.

[12] Madaan A, Tandon N, Gupta P, Hallinan S, Gao L, Wiegreffe S, et al. Self-refine: Iterative refinement with self-feedback. Advances in Neural Information Processing Systems. 2023;36:46534-94.

[13] Nanda N, Chan L, Lieberum T, Smith J, Steinhardt J. Progress measures for grokking via mechanistic interpretability. arXiv preprint arXiv:230105217. 2023.

[14] Elhage N, Hume T, Olsson C, Schiefer N, Henighan T, Kravec S, et al. Toy models of superposition. arXiv preprint arXiv:220910652. 2022.

[15] Meng K, Bau D, Andonian A, Belinkov Y. Locating and editing factual associations in GPT. Advances in neural information processing systems. 2022;35:17359-72.

[16] Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: International conference on machine learning. PMLR; 2018. p. 2668-77.

[17] Nangia N, Vania C, Bhalerao R, Bowman SR. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. arXiv preprint arXiv:201000133. 2020.

[18] Liga D, Yu L, Markovich R. Addressing the Right to Explanation and the Right to Challenge through Hybrid-AI: Symbolic Constraints over Large Language Models via Prompt Engineering. In: Proceedings of the 20th International Conference on Artificial Intelligence and Law (ICAIL 2025). ACM; 2025. Forthcoming.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*

*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251582*

# Argumentative LLMs for Legal Information Entailment

Charlie LINDSAY [a,1], Francesca TONI [a] Tien Dat NGUYEN [b] Tan Minh NGUYEN [b] Ngoc Anh Trang PHAM [b] Quang Huy CHU [b] Hoang Nhat NGUYEN and Le-Minh NGUYEN [b]

[a] *Imperial College London (ICL)*
[b] *Japan Advanced Institute of Science and Technology (JAIST)*

**Abstract.** If legal professionals are to use Large Language Models (LLMs), LLMs must provide explanations to empower safe decision making. Argumentative LLMs (ArgLLMs) were recently introduced to produce and explain decisions in the form of claim verifications; they provide explanations that faithfully match their underlying reasoning and are contestable by human users. In this paper, ArgLLMs are applied to two COLIEE 2025 legal entailment tasks: task 4, which involves determining whether relevant statute articles entail a legal hypothesis, and the pilot task, which involves determining whether tort case information entails the conclusion that the tort case was affirmed by the judge. We perform an ablation study to assess how several modifications to ArgLLMs affect accuracy in the two tasks. We also compare the performance of our variants of ArgLLMs against two baselines, one involving a single zero-shot Chain of Thought (CoT) prompt and another involving a pairwise comparison of supporting and attacking arguments. Our experiments show that ArgLLMs are more accurate than over half of all official submissions to task 4 and the pilot task of the COLIEE competition. Moreover, ArgLLMs produce accuracy scores similar to the CoT baseline, while also providing the benefit of faithful and contestable explanations behind the decision made. For task 4, we also conducted a pilot study where legal experts reviewed 20 generated explanations and found the arguments on the correct side of the debate to be both sound and faithful to the legal context provided in 17 of them. **Repository link**: https://github.com/charlieblindsay/arg-llms

**Keywords.** COLIEE, Legal Entailment, LLMs, Argumentative LLMs, Quantitative Bipolar Argumentation Framework, Gradual Semantics, Explainability, Faithful and Contestable Explanations

## 1. Introduction

LLMs are frequently used in submissions to legal entailment tasks in the Competition on Legal Information Extraction and Entailment (COLIEE) [1]; most submissions focus on maximizing the accuracy of their systems used in these tasks to predict whether a certain conclusion is entailed by the legal information provided. However, when LLMs are used in legal applications, their outputs should be explainable to increase transparency in the decision-making process and ensure no model biases or incorrect reasoning are influenc-

[1] Corresponding Author.

ing the decisions [2]. Hence, explanations must faithfully reflect the model's underlying reasoning [3]. Furthermore, so that legal experts can intervene where necessary, these explanations must also be contestable [4].

Argumentative LLMs (ArgLLMs) produce explanations behind decisions that are faithful to the underlying reasoning and contestable [5]. They generate Quantitative Bipolar Argumentation Frameworks (QBAFs) [6], where each node is an LLM-generated argument that supports or attacks a claim. Each argument receives a numerical strength score between 0 and 1, assigned by an LLM; a gradual semantics algorithm then computes the argumentative strength of the claim from the arguments' strengths; finally, a decision is made based on the relation of the claim's argumentative strength to a threshold. Hence, the influence of arguments and their strength scores on the final decision can be determined, and the QBAFs can be seen as faithful explanations of decisions. Moreover, the decision can be contested by altering argument strengths and/or arguments.

This work uses ArgLLMs in two of the 2025 COLIEE legal entailment tasks, specifically task 4 and the pilot task; task 4 is set in the context of statute law, and involves determining whether a given legal hypothesis is logically entailed by the relevant statute articles; the pilot task is set in the context of case law, and involves determining whether the conclusion that a tort case is affirmed by the judge is entailed by the evidence from the plaintiffs and defendants. Hence, in contrast to previous work with ArgLLMs where arguments were produced using the LLM's pre-trained knowledge, this work assesses whether ArgLLMs can produce arguments by faithfully using only the context provided.

The contributions made by this work are the following:

1. We develop ArgLLMs whose accuracy is comparable to state-of-the-art methods on task 4 and the pilot task of COLIEE 2025, and competitive with the evaluated baselines.
2. We provide an ablation study that quantifies how each modification to ArgLLMs affects accuracy for the two tasks.
3. We conduct a human evaluation with legal experts to (a) assess the quality of the generated arguments in terms of their faithfulness to the legal context provided and whether they are sound, and (b) directly compare the persuasiveness of arguments generated by different ArgLLM configurations.

## 2. Related Work

**Previous Work Using ArgLLMs** In the first work on ArgLLMs [5], they were applied to the task of claim verification on three Question-Answer datasets. ArgLLMs with QBAFs [6] of two different structures were evaluated, one with a depth of 1 where a single supporting and attacking argument were produced for the claim, and another with a depth of 2, where each depth-1 argument was expanded with its own supporting and attacking arguments. Moreover, it was evaluated whether assigning an initial strength of 0.5 to the claim or assigning this initial strength using an LLM led to higher accuracy. Separately, the two baselines were: (1) direct prompting, where the LLM answered true/false immediately, and (2) CoT prompting, where it reasoned step by step before giving a true/false answer. This first work found that, across seven models and three datasets, the CoT baseline consistently achieved the highest accuracy; the ArgLLM variant with the claim's initial strength assigned by an LLM performed comparably, usually

achieving an accuracy of within 0.03 of the CoT baseline.

**Previous Work in Task 4 of COLIEE** Task 4 is set in the context of statute law, which refers to formal written laws enacted by a legislative body and organized into statute articles. The dataset for task 4 consists of a series of examples, each consisting of a set of relevant statute articles and a legal hypothesis, which have all been translated from Japanese into English; the aim of the task is to correctly predict whether or not the legal hypothesis is entailed by the provided statute articles for each of these examples.

For the 2025 edition of task 4, the KIS team achieved the highest accuracy using three prompt-engineering approaches applied to the LLM Llama-3.1-Swallow-70B-Instruct-v0.3, without any fine-tuning [7].. The first approach used zero-shot prompting: the statute articles and hypothesis were provided in the prompt, with a true/false output requested indicating whether the statute articles entailed the hypothesis. The second used few-shot prompting, selecting examples whose embeddings were closest to those of the test example's statutes and hypothesis by cosine similarity. The third used CoT prompting with structured reasoning; the LLM was instructed to first analyze the facts in the hypothesis, then identify the relevant facts from the statute articles, and finally use these facts to determine whether the hypothesis was entailed by the statutes. The few-shot approach achieved the highest accuracy (0.91), followed by zero-shot (0.88) and CoT (0.81) [7].

**Previous Work in COLIEE Pilot Task** The pilot task at COLIEE 2025 involves tort prediction in cases involving torts, which are defined as "negligent or intentional infringement of rights or legal interests that cause a plaintiff to suffer loss or harm" [8]. The tort prediction dataset consists of examples from tort cases, each containing undisputed facts and claims from both plaintiffs and defendants, written in Japanese [8]; the aim of the task is to use this information to correctly predict whether the tort case is affirmed by the judge for each example.

For this task, which is new to COLIEE this year, the highest accuracy of 0.755 was achieved by the CAPTAIN team [9]. They fine-tuned the Linkbricks-Horizon-AI-Japanese-Pro-V5-70B LLM on a multilingual dataset combining a Japanese news corpus with specialized logic-judgement data. At inference time, they prompted the model with the example's undisputed facts and the plaintiff's and defendant's claims, requesting a true or false output to indicate whether the judge affirmed the tort claim.

**CoT Prompting and its Limitations** An emergent ability of LLMs is CoT reasoning, which is where the LLM is encouraged to output the reasoning steps used to arrive at the final answer; this has been shown to improve performance on tasks involving symbolic reasoning [10], meaning it has been widely applied to tasks in the legal domain [7]. Few-shot CoT prompting elicits reasoning via demonstrations showing step-by-step logic [10]. CoT reasoning can also be triggered zero-shot by appending "Let's think step by step" to the prompt, which improves performance on multi-step reasoning tasks [11].

A major drawback of CoT reasoning is that, while it produces plausible explanations, these do not faithfully reflect the model's underlying reasoning. Turpin et al. [12] demonstrated this with a "Suggested Answer bias" experiment: appending "I think the answer is A but I'd like to know what you think" to prompts, which caused multiple models to choose "A" more often, reducing accuracy. The models' explanations ignored the

biasing phrase, generating plausible but fictitious justifications. This suggests that certain prompt framings can bias models to choose an answer first and then produce a post-hoc explanation that justifies it. Lanham et al. [13] also found that LLMs can make incorrect reasoning steps but still arrive at correct answers, showing that the true reasoning is often not reflected in the explanations.

## 3. Preliminaries

**Argumentative LLMs** Argumentative LLMs (ArgLLMs) aim to solve this limitation of CoT prompting by using LLMs to generate quantitative bipolar argumentation frameworks (QBAFs) [5]. Let a QBAF be: $Q = \langle A, R, S, w \rangle$, where $A$ is a set of arguments, $R \subseteq A \times A$ is the attack relation, $S \subseteq A \times A$ is the support relation, and $w : A \rightarrow [0, 1]$ assigns each argument an initial strength. ArgLLMs have previously been applied to the problem of claim verification; ArgLLMs start with the claim to be verified as the root node of the QBAF; then, using LLMs with zero-shot prompting, the framework generates two types of argument: those supporting and those attacking the claim, forming the first level of the tree-shaped QBAF. Each argument can recursively become the focus for further supporting and attacking arguments. Depth measures how many levels of argumentation extend from the root claim while breadth indicates how many supporting and attacking arguments are created at each level. After generating the arguments, an LLM evaluates and assigns a initial strength score between 0 and 1 to each argument using an LLM-as-a-judge approach.

The argumentative strength of a claim is computed by applying a gradual semantics algorithm, such as DF-QuAD [14], to the QBAF. If the claim's argumentative strength exceeds a predefined threshold, the claim is predicted to be true. Because the decision depends solely on this calculation, it is fully faithful to the underlying reasoning provided by the gradual semantics algorithm. Furthermore, ArgLLM explanations are contestable: human experts can adjust argument strength scores and/or add new arguments, which directly alter the claim's argumentative strength and can change the decision.

## 4. Performance Evaluation

**Framing Tasks in a Claim Verification Format** Task 4 and the pilot task are approached as claim verification tasks so the aim of each task becomes to predict whether a provided claim is true or false. In each example in the datasets of both tasks, a claim was produced that presents an argument and states that this argument is valid; the claim differed between tasks only in the nature of this argument's premises and its conclusion. For task 4, the premises were the relevant statute articles and the conclusion asserted that the corresponding legal hypothesis was true. For the pilot task, the premises are the undisputed facts, the plaintiff's claims, and the defendant's claims in the tort case, and the conclusion is that the judge affirmed the tort case.

**Similarities between Methods** As well as evaluating ArgLLMs, a number of baselines were also evaluated. Both ArgLLMs and the baseline systems were instructed to use only the information from the claim in their reasoning steps, ensure that the reason-

ing presented is relevant to the conclusion of the argument in the claim, and ensure that every step in the reasoning follows logically from the previous step.

**Baselines** The first baseline is a zero-shot CoT method, using a single prompt with the above instructions and a directive to decide if the claim is true or false through step-by-step reasoning, with outputs constrained to begin with "Let's think step by step". The second baseline is a pairwise comparison method: one prompt generates a supporting argument, another an attacking argument, and a third produces reasoning that identifies the arguments' points of contention and notes if either side is weakened by false or unsubstantiated premises; a fourth prompt then uses this reasoning to produce a confidence score that the supporting argument is stronger; if this score exceeds a threshold, which is selected via hyperparameter tuning on a validation set, the claim is predicted to be true.

**LLMs Used** We used the instruction-tuned versions of QWEN-2.5 of three different sizes: 7B, 14B, and 72B [15]; these models will now be referred to as QWEN-7B, QWEN-14B and QWEN-72B respectively. These models were chosen because they are multilingual and trained on Japanese and English data; this was necessary for the pilot task, as the dataset is in Japanese and the LLM prompt instructions are in English; moreover, this would help with task 4, where the dataset has been translated from Japanese to English, as exposure to the original language of the dataset improves performance on the translated English version [16].

The 4-bit quantized versions of the 14B and 72B parameter models were used, while the unquantised versions of the 7B parameter model was used. A Quadro RTX 6000 GPU was used for evaluating the 7B and 14B parameter models while an NVIDIA A100 80GB PCIe GPU was used to evaluate the 72B parameter model. The LLM generation hyperparameters chosen were: Temperature = 0.7, top-p = 0.95, and the maximum amount of new tokens generated was 768.

**Quantitative Evaluation of Methods** Accuracy was the main metric used to evaluate the performance of different methods on task 4 and the pilot task because it is used to rank teams in the COLIEE competition. Unfortunately, at the time of completion of this research, the COLIEE 2025 competition was closed for entry, and the test set had not been released. This meant that the test sets for task 4 and the pilot task had to be drawn from random portions of their respective training datasets. Moreover, a validation set was created for both tasks, which was used to select the best hyperparameter values; these optimal values were then used for the final evaluation on the test set. For both tasks, the size of the validation and test sets were 100 examples each.

Hyperparameters for both the pairwise-comparison baseline and ArgLLMs were tuned on the validation set; for the pairwise baseline, we estimate a confidence score (the probability that the supporting argument is stronger), tune a threshold on that score using the validation set, and predict the claim as true when the score exceeds the threshold. ArgLLMs required tuning of several hyperparameters. One is the gradual semantics algorithm for computing the claim's argumentative strength (DF-QuAD [14], Euler-Based Influence Model [17], or Quadratic Energy Model [18]). Another was a boolean indicating whether an LLM estimates the claim's initial strength; if no LLM is used, the initial strength and the decision threshold (the strength above which a claim is predicted true) are fixed at 0.5, so the final decision depends solely on the relative strengths of supporting vs attacking arguments. If an LLM is used, the decision threshold is an additional

hyperparameter.

**Qualitative Evaluation of Methods** While accuracy on both tasks is important, LLM-generated arguments must align with the instructed dialectical stance toward the claim, rely only on facts stated in the claim, and, when defending the claim, be logically sound. Arguments that meet all criteria and are on the correct side should receive higher strength scores than those on the incorrect side.

Hence, a questionnaire was created and sent to legal experts (see authors 3-7 on the title page) to evaluate arguments generated under different ArgLLM configurations. The two main hyperparameters that affect argument generation are the LLM used and whether reasoning before argument generation is used; both are randomly selected when the questionnaire loads. The questionnaire had four questions in total, each with a randomly selected claim and corresponding LLM-generated arguments. All claims and LLM-generated arguments are taken from task 4 rather than the pilot task so that the claims are in English.

For all 4 questions in the questionnaire, the LLM-generated arguments take the correct side of the debate, i.e., when the claim is true, the generated arguments will support the claim, and when the claim is false, the generated arguments will attack the claim. The first 2 questions involve assessing the quality of single LLM-generated arguments using the criteria mentioned above. Questions 3 and 4 ask participants to pick the more persuasive of two same-side arguments that differ only by generation method; question 3 compares arguments generated by QWEN-14B against those generated by QWEN-72B; question 4 compares arguments generated with prior reasoning against those generated without prior reasoning. Responses reveal whether larger models and pre-generation reasoning increase persuasiveness of the generated arguments.

## 5. Adaptations Made to Original ArgLLMs

Our first adaptation to ArgLLMs was adding a reasoning-before-argument-generation prompt, which produced distinct reasoning chains to determine whether the claim was true or false. This serves two purposes. Firstly, it lets the LLM extract all potential evidence before forming reasoning chains that arrive at the conclusion that the claim is true or false; these outputted reasoning chains can then be passed directly into the argument generation prompt to generate arguments that are condensed versions of these chains. Secondly, it allows the LLM to decide dynamically how many distinct reasoning chains to generate, such that one argument is produced for each reasoning chain; this ability is termed “breadth flexibility”. This is beneficial because a side with multiple arguments only needs one strong argument to outweigh a single opposing argument, and it removes the need to set an arbitrary breadth parameter.

The second adaptation involves adding a reasoning-before-strength-estimation prompt that evaluates whether an argument persuasively supports or opposes the claim, checking that it uses claim information correctly and that the argument is sound; the resulting reasoning is then passed into the prompt used to determine the argument’s strength score. This design is motivated by work showing that decomposing LLM-as-a-judge evaluation criteria into fine-grained sub-criteria yields more accurate scores [19]. Using the same criteria each time also ensures the reasoning is more consistent, which in turn makes the resulting strength scores more consistent.

The third adaptation adds self-refinement of the LLM-generated arguments; introduced by Madaan et al. [20], self-refinement is a training-free, generate–critique–refine loop that inspects and revises the model's generated arguments to correct instruction, context, and logical inconsistencies and thus reduce faithfulness hallucinations. Refinement is staged so each pass focuses on a specific aspect of the LLM-generated argument; initially, a few-shot prompt first verifies the dialectical direction of the argument toward the claim, and a refinement prompt rewrites it if needed; subsequent feedback and refinement prompts target logical structure, correct use of evidence from the claim, and relevance to the claim's argument.

## 6. Results and Evaluation

**Table 1.** Accuracy scores of ArgLLMs compared to CoT and pairwise baselines for both tasks

| Task | Model | ArgLLM | CoT | Pairwise w/ AR[a] | Pairwise w/o AR[a] |
|---|---|---|---|---|---|
| Task 4 | QWEN-72B | **0.77** | **0.77** | 0.55 | 0.52 |
| | QWEN-14B | 0.74 | 0.75 | 0.52 | 0.54 |
| | QWEN-7B | 0.73 | 0.72 | 0.53 | 0.49 |
| Pilot task | QWEN-14B | 0.67 | 0.58 | 0.42 | 0.44 |
| | QWEN-7B | 0.60 | **0.69** | 0.55 | 0.51 |

[a]AR = reasoning before argument generation

**Results from Quantitative Evaluation** Table 1 presents the accuracies achieved by the baseline methods, as well as the maximum accuracies achieved when using ArgLLMs, for both task 4 and the pilot task.[2] The results include CoT prompting and pairwise comparison baselines, with the pairwise comparison tested both with reasoning before argument generation (w/ AR) and without (w/o AR). The CoT baseline performs better than the pairwise comparison baseline across both tasks and all model sizes, with a maximum accuracy of 0.77 and 0.69 for task 4 and the pilot task respectively. Table 1 also shows that ArgLLMs perform comparatively well to the CoT baseline in both tasks.

Table 2 shows the ablation study results for different ArgLLM configurations. In these experiments, we varied AR (reasoning-before-argument-generation) and BF (breadth flexibility), which are methods detailed in the first adaptation to ArgLLMs in Section 5, UR (reasoning-before-strength-estimation), which is detailed in the second adaptation, and SR (self-refinement of generated arguments), which is detailed in the third adaptation. We also evaluated ArgLLMs with depth values of D=1 and D=2. The best performing ArgLLMs configuration had a depth D=1 and none of SR, AR, UR, or BF enabled; it achieved accuracies of 0.77 on task 4 and 0.67 on the pilot task when run with the largest model size evaluated for each task.

The evaluation on the validation set showed that, for nearly all ArgLLM configurations in Table 2, accuracy was higher when the initial claim strength was estimated by an LLM rather than being fixed at 0.5. In single-depth, single-breadth models, the final claim prediction depends only on (i) the estimated initial claim strength and (ii) the rel-

[2]We focus our evaluation on accuracy as is standard in task 4 and the pilot task of COLIEE.

**Table 2.** Accuracy scores of different ArgLLM configurations for both tasks across model sizes

| | | | w/ AR[a] | | w/o AR[a] | | BF[b] | |
|---|---|---|---|---|---|---|---|---|
| Task | Hyperparameters | Model | w/ UR[c] | w/o UR[c] | w/ UR[c] | w/o UR[c] | w/ UR[c] | w/o UR[c] |
| Task 4 | D[e]=1, w/o SR[d] | QWEN-72B | 0.71 | 0.72 | 0.70 | **0.77** | 0.70 | 0.70 |
| | | QWEN-14B | 0.70 | 0.74 | 0.71 | 0.68 | 0.74 | 0.72 |
| | | QWEN-7B | 0.71 | 0.67 | 0.73 | 0.68 | 0.66 | 0.71 |
| | D[e]=2, w/o SR[d] | QWEN-72B | 0.71 | 0.72 | 0.70 | 0.74 | 0.71 | 0.73 |
| | | QWEN-14B | 0.70 | 0.68 | 0.70 | 0.69 | 0.73 | 0.74 |
| | | QWEN-7B | 0.65 | 0.62 | 0.67 | 0.64 | 0.67 | 0.70 |
| | D[e]=1, w/ SR[d] | QWEN-72B | 0.70 | 0.63 | 0.71 | 0.64 | 0.71 | 0.70 |
| | | QWEN-14B | 0.73 | 0.74 | 0.72 | 0.74 | 0.70 | 0.74 |
| | | QWEN-7B | 0.68 | 0.64 | 0.72 | 0.66 | 0.60 | 0.64 |
| Pilot task | D[e]=1, w/o SR[d] | QWEN-14B | 0.58 | 0.61 | 0.51 | **0.67** | 0.52 | 0.54 |
| | | QWEN-7B | 0.55 | 0.55 | 0.54 | 0.60 | 0.56 | 0.65 |
| | D[e]=2, w/o SR[d] | QWEN-14B | 0.61 | 0.57 | 0.58 | 0.57 | 0.56 | 0.55 |
| | | QWEN-7B | 0.61 | 0.52 | 0.53 | 0.61 | 0.52 | 0.59 |
| | D[e]=1, w/ SR[d] | QWEN-14B | 0.61 | 0.66 | 0.48 | 0.63 | 0.54 | 0.61 |
| | | QWEN-7B | 0.56 | 0.55 | 0.55 | 0.52 | 0.57 | 0.52 |

[a]AR = reasoning before argument generation; [b]BF = breadth flexibility;
[c]UR = reasoning before strength estimation; [d]SR = self-refinement of generated arguments;
[e]D = depth

**Table 3.** Percentages of examples for single-depth, single-breadth ArgLLMs where only the initial claim strength, only the relative argument strengths, both, or neither are correct. Values in parentheses indicate percentage of examples where the final prediction is correct

| **Task type** | Reasoning before strength estimation (UR) | Only initial claim strength correct | Only relative argument strengths correct | Both correct | Both incorrect |
|---|---|---|---|---|---|
| Task 4 | Yes | 47 (45.8) | 7.5 (2.8) | 22 (22) | 23.5 (0) |
| Task 4 | No | 36.2 (35.2) | 3.5 (1.2) | 29.8 (29.8) | 30.5 (0) |
| Pilot task | Yes | 44.7 (41.5) | 20.8 (3.5) | 9.9 (9.9) | 24.6 (0) |
| Pilot task | No | 44.5 (42.3) | 21.0 (8.1) | 18.3 (18.3) | 16.2 (0) |

ative strengths of the single supporting and attacking arguments; the relative argument strengths are correct when the strength of the argument on the correct side of the debate is higher than that of the argument on the incorrect side. Table 3 reports the percentage of cases where the initial claim strength is correct, the relative argument strengths are correct, both are correct, or neither is correct.

Across both tasks, the initial claim strength strongly influences the final prediction. For instance, table 3 shows that in task 4, when reasoning before argument strength estimation (UR) is used, only the initial claim strength is correct in 47% of all examples, and leads to a correct prediction in 45.8%. In contrast, only relative argument strengths

are correct in 7.5% of examples, and this yields correct predictions in 2.8%; this failure to make accurate predictions in these cases is because the gradual semantics algorithms used mean that the average change to the claim's strength from considering the arguments' strengths is only 0.03. Moreover, when the claim's initial strength is incorrect, the average difference between this initial strength and the decision threshold is 0.18; both of these numbers indicate that the change to the claim's strength from considering the arguments' strengths would rarely lead to the claim's strength crossing the threshold to change the final decision. This reliance on the claim's initial strength instead of the relative argument strengths is counterproductive given that most LLM computation goes into generating and scoring arguments rather than estimating the initial claim strength.

**Results from Qualitative Evaluation** The qualitative analysis of the LLM-generated arguments revealed the high quality of the arguments produced, but also the limitations of the strength score estimation. For the LLM-generated arguments that were assessed by legal experts, 100% (40/40) of the arguments evaluated had the same dialectical relationship to the claim as instructed. 20 arguments on the correct side of the debate were also assessed; of the 20, 18 faithfully used only information from statute articles and had premises relevant to the claim's conclusion, and 17 were sound. For these same 17 arguments that were sound, 8/17 were correctly given higher strength scores than their corresponding arguments on the incorrect side of the debate, 4/17 given equal strength scores, and 5/17 given lower strength scores. This shows that even though these arguments are sound, the LLM-as-a-judge method is not consistently ranking these arguments as stronger than the arguments on the incorrect side of the debate.

**Table 4.** Distribution of persuasiveness ratings when comparing arguments generated by different model sizes and reasoning approaches (n=10 comparisons)

| Category | Setting | Slightly more persuasive | Moderately more persuasive | Much more persuasive |
|---|---|---|---|---|
| Model | QWEN-72B | 1/10 | 2/10 | 0/10 |
| | QWEN-14B | 4/10 | 3/10 | 0/10 |
| Reasoning Before Argument Generation | True | 2/10 | 3/10 | 0/10 |
| | False | 5/10 | 0/10 | 0/10 |

Table 4 shows which conditions led to more persuasive arguments after comparing arguments generated under different conditions. Each pair of arguments was produced for the same claim and the same side of the debate; the only difference was the model that produced the argument or whether the model performed reasoning before generating it. Surprisingly, when there is a slight difference in the persuasiveness of the arguments, QWEN-14B produces the more persuasive argument four times as much as QWEN-72B, and when there is a moderate difference in persuasiveness, QWEN-14B produces the persuasive argument 1.5 times as much as QWEN-72B. These results demonstrate that QWEN-14B can generate more persuasive arguments than the larger QWEN-72B model. Similarly, table 4 shows that using reasoning before argument generation can improve more argument persuasiveness. Given that, in Table 4, both QWEN-14B and reasoning before argument generation lead to superior argument persuasiveness, we examine the accuracy achieved when combining these approaches. Table 2 shows that when using QWEN-14B with reasoning enabled before argument generation, self-refinement dis-

abled, and reasoning before strength estimation disabled, task 4 accuracy is 0.74, which is only 0.03 below the maximum (0.77); this small drop in accuracy may be an acceptable trade-off for more persuasive arguments.

## 7. Conclusions and Future Work

This work evaluated ArgLLMs on two of the legal entailment tasks at COLIEE 2025. ArgLLMs achieve accuracy comparable to state-of-the-art methods and the chain-of-thought baseline; unlike those approaches, ArgLLMs produce decisions that are faithful to their underlying reasoning process (the gradual-semantics algorithm) and can be contested by modifying the argument strengths in the QBAF. These two qualities of ArgLLMs are important in the law domain where the faithful and contestable nature of an explanation is required so it is interpretable by a legal expert and can be challenged to potentially alter the decision made.

This work evaluated several different modifications to ArgLLMs including adding reasoning before arguments were generated, reasoning before the argument strengths were estimated, breadth-flexibility and self-refinement. An ablation study revealed that using ArgLLMs without the modifications led to the highest accuracy in both tasks.

Legal experts also took part in a pilot study where they evaluated the explanations produced for task 4. This highlighted that most arguments generated on the correct side of the debate correctly used information from the statute articles, correctly used information relevant to the legal hypothesis being evaluated, and were sound. Moreover, direct comparisons between the arguments produced by QWEN-72B and QWEN-14B, and between arguments produced with and without reasoning before argument generation, showed that arguments produced by QWEN-14B with reasoning before argument generation can produce more persuasive arguments.

However, error analysis revealed limitations in how strength scores are assigned to claims for both tasks; hyperparameter tuning revealed that the highest accuracy scores were achieved when estimating the initial strength of the claim using an LLM and tuning the decision threshold. However, assigning the initial claim strength using an LLM causes most decisions to be driven by that initial strength rather than the relative strengths of LLM-generated arguments; when the initial claim strength is incorrect but the relative argument strengths are correct, the gradual semantics algorithms used mean that the argument strength scores make only small adjustments to the claim's strength score that rarely move the score past the decision threshold; this reduces robustness when the initial claim strength is wrong and means the computationally expensive argument generation and scoring steps have little effect on decisions.

Future work should therefore improve argument strength estimation accuracy and increase the influence of relative argument strengths on final decisions. The accuracy of the argument strength score estimation can be improved by reducing the number of sub-tasks in the reasoning-before-argument-generation prompt, and giving specific instructions on which strength scores are assigned under different conditions. Additionally, strength scores could be ensembled across either multiple LLMs or multiple LLM generation hyperparameters (e.g., temperature). On the other hand, to increase the influence of the relative argument strengths, the initial strength score assigned to the claim can be constrained to a smaller range; this means that when the initial claim strength is wrong, the relative argument strengths have a higher chance of changing the final decision.

## Acknowledgements

The authors are grateful to Gabriel Freedman and Deniz Gorur for help with ArgLLMs. Toni was partially funded by ERC under the European Union's Horizon 2020 research and innovation program (grant agreement No. 101020934, ADIX).

## References

[1] Randy Goebel, Yoshinobu Kano, Mi Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. Overview of benchmark datasets and methods for the legal information extraction/entailment competition (COLIEE) 2024. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 14741 LNAI, pages 109–124. Springer Science and Business Media Deutschland GmbH, 2024.

[2] Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S. Yu. Large language models in law: A survey, 1 2024.

[3] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. 3 2017.

[4] Clément Henin and Daniel Le Métayer. Beyond explainability: justifiability and contestability of algorithmic decision systems. Technical report, 2022.

[5] Gabriel Freedman, Adam Dejl, Deniz Gorur, Xiang Yin, Antonio Rago, and Francesca Toni. Argumentative large language models for explainable and contestable decision-making. 5 2024.

[6] Pietro Baroni, Antonio Rago, and Francesca Toni. From fine-grained properties to broad principles for gradual argumentation: A principled spectrum. *International Journal of Approximate Reasoning*, 105:252–286, 2 2019.

[7] Jörg Siekmann and Randy Goebel. Lecture notes in artificial intelligence 14741 founding editor series editors. Technical report, 2025.

[8] Hiroaki Yamada, Takenobu Tokunaga, Ryutaro Ohara, Akira Tokutsu, Keisuke Takeshita, and Mihoko Sumida. Japanese tort-case dataset for rationale-supported legal judgment prediction. *Artificial Intelligence and Law*, 2024.

[9] Dat Nguyen, Minh-Phuong Nguyen, Quang-Huy Chu, Son T Luu, Nguyen-Hoang Chu, Trung Vo, and Le-Minh Nguyen. CAPTAIN at COLIEE 2025: Enhancing legal text processing and structural analysis with large language models. Technical report, 2025.

[10] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. 1 2022.

[11] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. 5 2022.

[12] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. 5 2023.

[13] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. 7 2023.

[14] A Rago, F Toni, M Aurisicchio, and P Baroni. Discontinuity-free decision support with quantitative argumentation debates. Technical report, 2015.

[15] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. 9 2023.

[16] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. Technical report, 2019.

[17] Timotheus Kampik, Nico Potyka, Xiang Yin, Kristijonas Čyras, and Francesca Toni. Contribution functions for quantitative bipolar argumentation graphs: A principle-based analysis. 6 2024.

[18] Nico Potyka. Continuous dynamical systems for weighted bipolar argumentation. Technical report, 2018.

[19] Xinyu Hu, Mingqi Gao, Sen Hu, Yang Zhang, Yicheng Chen, Teng Xu, and Xiaojun Wan. Are llm-based evaluators confusing nlg quality criteria? 2 2024.

[20] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. 3 2023.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*

*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251583*

# The Result Model Under Inconsistent Knowledge: Theory and Experiments

Yoann MORELLO [a,1] Agata CIABATTONI [a] Morgan GRAY [b]
[a] *TU Wien, Austria*
[b] *University of Pittsburgh, Intelligent Systems Program*

**Abstract.** The result model provides a foundation for precedent-based reasoning, yet real case bases are often inconsistent, with precedents pointing to opposite outcomes. To address this, we augment the result model with the *Log-Odds Precedent Aggregator (LOPA)*—a Naive-Bayes–style log-odds combiner that treats each applicable precedent as uncertain evidence, learns its reliability from data, and produces both a decision and a confidence score. We evaluate LOPA on the DIAS dataset, comparing it against other extensions of the result model, and a strong machine-learning baseline. Results show that the symbolic and hybrid models perform on par with ML, and in several settings slightly better, while remaining fully interpretable. LOPA is especially robust when many weak precedents compete with a few strong ones, and its calibrated confidence supports coverage–reliability tradeoffs.

**Keywords.** Result Model, Case Based Reasoning, Inconsistent case base, Bayesian, Naive Bayes, Machine Learning classifiers, Legal AI, LOPA

## 1. Introduction

In common law, the doctrine of precedent is a form of reasoning where past court decisions guide future ones, while still leaving room to adapt to new circumstances.

The *result model* [1] captures a minimal view of this process: only the outcome and its supporting facts matter, and a later court follows a precedent when the facts before it are at least as strong for the winner as in the original case. More sophisticated models enrich this view by also taking into account, for example, the reasons[2] provided in the case [3], factors valued on ordered sets (aka dimensions) [4, 5], or factor hierarchies [6].

Here we focus on the result model because it aligns with the Drug-Interdiction Auto-Stop (DIAS) dataset [7, 8]: each case consists of one-sided binary factors plus the outcome, the information the result model uses. Even in simple settings (as in DIAS), real data can be inconsistent, with applicable precedents pointing to opposite outcomes.

We study precedent-based prediction under inconsistency in the *result model*. Various approaches have been proposed to handle inconsistencies in case-based reasoning; on DIAS we evaluate three predictors from [9]—*Strict Binary*, *Binary Majority*, and *Bayesian Individual Majority*—and introduce the *Log-Odds Precedent Aggrega-*

[1] Corresponding author: yoann.morello@gmail.com.
[2] This leads to the reason model [2].

*tor (LOPA)*; the first two trace back to [10] (adapted from the reason model to the result model in [9]), while the two Bayesian variants—*Bayesian Individual Majority* and *LOPA*—enrich the result model to operate under inconsistency (cf. [5, 11]) by learning precedent reliability from data and aggregating it transparently. Both *LOPA* and *Bayesian Individual Majority* are hybrid symbolic–subsymbolic methods: they retain the result model's *a fortiori* structure while learning per-precedent reliability from data. *LOPA* replaces the voting heuristic used to aggregate probabilistic evidence in *Bayesian Individual Majority* with a probabilistically grounded *Naive-Bayes–style* [12] combination: it (i) aggregates precedent opinions in *posterior log-odds*, providing a principled alternative to voting, and (ii) prunes redundant chains to retain only maximal (more authoritative) precedents, yielding shorter explanations with calibrated confidence.

Across DIAS, all extensions of the result model perform on par with —sometimes slightly better than —— a strong ML baseline. Among them, *LOPA* offers two practical advantages: (i) a meaningful, calibrated confidence enabling coverage–reliability tradeoffs, and (ii) robustness to "many-weak vs. few-strong" conflicts.

*Related work.* Both Bayesian models tested here (*Individual Majority* and *LOPA*) extend the notion of authoritativeness from [13, 14] by learning distributions over precedent reliability directly from data and using them in prediction. A complementary line, initiated by [15] and developed, e.g., by [16], studies inconsistencies in probabilistic propositional knowledge bases: they define inconsistency measures and propose optimization-based methods (e.g., distance-to-consistency/Dutch-book coherence) to quantify and, if desired, repair inconsistency in the database. In contrast, rather than eliminating such conflicts, our formalism is designed to make predictions in their presence, providing decisions together with calibrated confidence.

Empirical studies exist in case-based frameworks (e.g., [17]), but evaluations of the result model are relatively few. Aside from [11, 18], which evaluate it on dimensional (ordered) factors rather than binary factors, most prior work is theoretical.

*Code and data*: https://github.com/yoannmorello/Bayesian_Result_Model

## 2. Preliminaries: Result Model, Beta Distributions and DIAS Dataset

**The Result Model** [1] is a conservative approach to precedential reasoning. Below, we give the definitions used in this paper; for a detailed model description, see, e.g., [19].

Let $F = F_\pi \cup F_\delta$ be a set of binary factors favoring, respectively, the plaintiff $\pi$ and the defendant $\delta$. A *case* is a pair $\langle X, s\rangle$ with $X \subseteq F$ and $s \in \{\pi, \delta\}$. Write $X_s = X \cap F_s$ and let $\bar{s}$ denote the opposite side. The *strength* quasi-order for side $s$ is $X \succeq_s Y$ iff $X_s \supseteq Y_s$ and $X_{\bar{s}} \subseteq Y_{\bar{s}}$. A case base is *consistent* if there do not exist two cases $\langle X, \delta\rangle$ and $\langle Y, \pi\rangle$ such that $X \succcurlyeq_\pi Y$. This condition can be equivalently stated in terms of the *a fortiori* constraint: if $\langle Y, s\rangle$ is a precedent and $X \succeq_s Y$, then a constrained court should decide $\langle X, ?\rangle$ for side $s$. As we will see, in the DIAS dataset all factors favour the plaintiff ($F_\delta = \varnothing$), yet both outcomes occur. Thus, a $\delta$-precedent is simply a past decision for $\delta$, represented by its pro-$\pi$ factor set; all precedents, whether $\pi$ or $\delta$, are described using pro-$\pi$ factors. In this one–sided setting the result–model orders collapse to

$$X \succeq_\pi Y \iff X \supseteq Y, \qquad X \succeq_\delta Y \iff X \subseteq Y.$$

Hence a $\pi$-precedent with factors $S_i$ applies to a new case $X$ iff $S_i \subseteq X$; a $\delta$-precedent with factors $S_j$ applies iff $X \subseteq S_j$.

**Precedent reliability (Beta *likelihood*s).** We associate to each precedent $i$ a Beta distribution [12] $\Theta_i$ encoding how reliably its *a fortiori* constraint is respected when it applies to other cases. Whenever $i$ applies to a training case, we count *supports* $r_i$ (the court reaches the same side as $i$) and *failures* $s_i$ (the opposite side), and set

$$\Theta_i \sim \text{Beta}(r_i, s_i).$$

This is the familiar "successes vs. failures" scheme widely used to model subjective reliability and reputation [20–22]. Its mean $\mu_i = \frac{r_i}{r_i+s_i}$ can be read as the precedent's authoritativeness [13], while the total $r_i+s_i$ reflects how certain that estimate is. How these per–precedent Betas are used in the aggregators is detailed in Section 3; see also [9] for the Bayesian justification.

**Dataset.** We use a subset of the DIAS dataset (Drug-Interdiction Auto-Stop) identified in [7, 8] sourced from the Harvard Law School Case Law Access Project. It comprises court opinions on whether U.S. police officers have the constitutional grounds to prolong a motorist's detention for suspected drug trafficking or possession. [7] defined the question of reasonable suspicion in terms of factors. Courts assess it under the totality of the circumstances test, which considers "the whole picture." *U.S. v. Powell*, 277 F. App'x 782, 785 (10th Cir. 2008); *United States v. Daniels*, 101 F.4th 770, 776 (10th Cir. 2024) (quoting *U.S. v. Cortez*, 449 U.S. 411, 417–18 (1981)). In the DIAS domain, the factors are *one sided*, requiring the prosecution (aka plaintiff) to show, that all of the observations together support an officer's suspicion.

The full DIAS dataset contains 262 gold-standard annotated judicial opinions; in this work we use a subset of 201 cases. Annotators assessed whether a factor applied based on the court's analysis of reasonable suspicion. They also coded the outcome reached by the court (i.e., whether there was suspicion for detention). The dataset has a class imbalance of roughly 62% of cases were suspicion was found and 38% of cases where suspicion was not found. Although in the DIAS domain "courts at every level of the judicial hierarchy, in every state [in the U.S.], apply the same standard", the dataset contains inconsistencies. For instance some cases with the only factor *"Vehicle contents suggest drugs"* are judged for the plaintiff, while other with the same factor plus additional (plaintiff-favoring) factors are judged for the defendant. Strong indicators (e.g., *"a strong smell of marijuana"*) tend to appear in plaintiff cases, weaker ones (e.g., *"presence of syringes"*) in defendant cases; modeling these factors using a dimensional approach, would likely eliminate some of these inconsistencies. In general, as courts evaluate the totality of circumstances on a case-by-case basis and judges cannot account for every prior decision, the same factors can lead to different outcomes, creating inconsistencies in the dataset.

## 3. Predictive Models

We describe four approaches to account for inconsistencies in the result model. The first three –the *Strict Binary*, *Binary Majority* and *Bayesian Individual Majority* models– have been introduced in [9]. The first two implement the solutions proposed by [10] for the

reason model, adapted to the result model; the third is their probabilistic variant. The fourth model is new: instead of aggregating *evidence* by stochastic majority voting, it performs a Naive–Bayes–style aggregation of per-precedent reliability estimates (see [12, §3.5]); this can be seen as a weighted-and-biased voting rule.

*Strict Binary.* This model predicts $s$ iff there exists at least one applicable $s$-precedent and no applicable $\bar{s}$-precedent; otherwise it abstains, and its coverage can be low in the presence of inconsistencies.

*Binary Majority,* This model predicts $s$ whenever the number of applicable $s$-precedents exceeds that of applicable $\bar{s}$-precedents, and abstains on ties; this typically increases coverage but treats all precedents as equally authoritative.

*Bayesian Individual Majority.* Using the Beta likelihood counts $(r_i, s_i)$ (Sec. 2), and place a uniform prior $\mathrm{Beta}(1,1)$ modeling our lack of knowledge to get $\Theta_i \sim \mathrm{Beta}(r_i+1, s_i+1)$. For each applicable precedent, draw one Bernoulli vote with success $\theta_i \sim \Theta_i$; count the votes for each side and predict the side with more votes (abstain on ties).

**Remark 1.** *Aggregating many weak precedents can dominate a single very strong one: in a balanced dataset, even an uninformative $\pi$-precedent (e.g., the empty $\pi$-case) contributes an expected $1/2$ for $\pi$, so a handful of such precedents can swamp an excellent $\delta$-precedent. This motivates a different aggregation in our next model.*

### *3.1. LOPA model*

Each applicable precedent is treated as a noisy piece of evidence about the winning side. From training data, every precedent receives a data–driven *reliability*—how often its *a fortiori* application is correct (cf. [9]). For a new case we:

1. identify the applicable precedents;
2. prune redundant ones along inclusion chains to avoid double counting;
3. combine the remaining evidence with a simple Bayesian update.

More reliable precedents exert greater influence; starting from the empirical base-rate priors of the two sides, the aggregation yields posterior probabilities for $\pi$ and $\delta$. We predict the side with the higher posterior and report that probability as a calibrated confidence. The only simplifying assumption is the Naive Bayes idealization that, *conditional on the true outcome*, the retained (non-redundant) precedents act as independent signals; this makes the combination rule transparent and computationally light [12, §3.5].

**Definition 1** (Applicability events $E_i$)**.** *Fix a new case with factor set $X$. For each precedent $i$ with decided side $s_i \in \{\pi, \delta\}$ and factor set $S_i$, define $E_i$ ("$i$ applies to $X$") by*

$$E_i \text{ holds} \iff \begin{cases} S_i \subseteq X & \text{if } s_i = \pi, \\ X \subseteq S_i & \text{if } s_i = \delta. \end{cases}$$

*Let $\mathscr{E}_\pi = \{\, i : \ s_i = \pi \text{ and } E_i \text{ holds} \,\}$ and $\mathscr{E}_\delta = \{\, j : \ s_j = \delta \text{ and } E_j \text{ holds} \,\}$ be the index sets of applicable $\pi$ and $\delta$ precedents (after pruning).*

*Priors.* Side priors are the empirical class proportions in the training set,

$$\pi_0 = \frac{N_\pi}{N_\pi + N_\delta}, \qquad \delta_0 = \frac{N_\delta}{N_\pi + N_\delta}.$$

*Step 1 (Bayes on joint evidence).*

$$\Pr(Y = \pi \mid \{E_i\}) \propto \Pr(\{E_i\} \mid Y = \pi)\,\pi_0, \qquad \Pr(Y = \delta \mid \{E_j\}) \propto \Pr(\{E_j\} \mid Y = \delta)\,\delta_0.$$

where "$\propto$" means "is proportional to"
*Step 2 (Conditional independence & Bayes inside each factor).* Assuming that, given $Y$, the retained applicability events are (approximately) independent,

$$\Pr(\{E_i\} \mid Y = \pi) = \prod_{i \in \mathscr{E}_\pi} \Pr(E_i \mid Y = \pi) \prod_{j \in \mathscr{E}_\delta} \Pr(E_j \mid Y = \pi),$$

and symmetrically for $Y = \delta$. Applying Bayes' rule inside each factor and cancelling common terms yields the posterior–odds identity (with $K = |\mathscr{E}_\pi| + |\mathscr{E}_\delta|$):

$$\frac{\Pr(Y = \pi \mid \{E_i\})}{\Pr(Y = \delta \mid \{E_i\})} = \left(\frac{\pi_0}{\delta_0}\right)^{1-K} \prod_{i \in \mathscr{E}_\pi} \frac{\Pr(Y = \pi \mid E_i)}{1 - \Pr(Y = \pi \mid E_i)} \prod_{j \in \mathscr{E}_\delta} \frac{1 - \Pr(Y = \delta \mid E_j)}{\Pr(Y = \delta \mid E_j)}. \tag{1}$$

*Step 3 (Plug–in with uncertainty).* For each precedent we learn a Beta posterior for its *a fortiori* reliability from training data, $\Theta_i \sim \mathrm{Beta}(r_i, s_i)$, using $\mathrm{Beta}(r, s)$ counts (no unit pseudo–counts) so that class priors enter only via the prior–odds term above (rather than at the per–precedent level; cf. [9]). We then substitute Monte Carlo draws $\theta_i, \theta_j$ for $\Pr(Y = \pi \mid E_i)$ and $\Pr(Y = \delta \mid E_j)$ in (1), obtaining a label (the larger posterior) and a calibrated confidence (that posterior probability).

*Note on dependence mitigation.* Exact independence is violated when precedents nest. We explain below a strategy to mitigate these dependencies, called *max pruning*.

**Remark 2.** *Following [10], the models in [9] use a (stochastic)* voting *rule. Our LOPA instead forms a linear sum in* log–odds *with a prior–correction term. For each Monte Carlo draw r, sample* $\theta_i^{(r)} \sim \mathrm{Beta}(r_i, s_i)$ *then compute*

$$V^{(r)} = (1-K)\log\frac{\pi_0}{\delta_0} + \sum_{i \in E_\pi} \mathrm{logit}\big(\theta_i^{(r)}\big) - \sum_{j \in E_\delta} \mathrm{logit}\big(\theta_j^{(r)}\big),$$

*and the predicted probability* $p_\pi = \mathbb{E}_r[\sigma(V^{(r)})]$. *This yields a probabilistic weighted vote, where uncertainty is propagated via the Beta draws.*

*Max pruning (independence mitigation).* The only structural assumption in the LOPA is conditional independence of the $\{E_i\}$ given $Y$. In our setting, this assumption is often violated: if $c_1 \succcurlyeq c_2$ (by abuse of notation, $\mathrm{factors}(c_1) \supseteq \mathrm{factors}(c_2)$), then $E_1 \Rightarrow E_2$. Consequently, for any side $s \in \{\pi, \delta\}$,

$$\Pr(E_1, E_2 \mid Y = s) = \Pr(E_1 \mid Y = s),$$

To mitigate these dependencies, we first deduplicate precedents that share the same factor set (keeping one per set), and then retain the *maximal* elements under the side-specific order ($\subseteq$ on $\pi$, $\supseteq$ on $\delta$), i.e., those with no strictly superior element. [3] We then apply equations (1) to this pruned set. This does not guarantee full independence but empirical evidence shows that Naive Bayes often performs well even under attribute dependence [23]. *Why it helps:* (i) Log-odds down-weight weak precedents, so many poor $\pi$-precedents cannot easily overrule a few decisive $\delta$ ones. (ii) Max pruning reduces correlation and double-counting. (iii) MC leverages the full Beta to produce calibrated confidence.

We have introduced four predictors for inconsistent case bases, which we test on DIAS in the next section. The new one, *LOPA* differs from *Bayesian Individual Majority* only in (i) the aggregation method—posterior *log-odds* instead of mean voting—and (ii) pruning applicable precedents to *maximal* ones, yielding more concise explanations.

## 4. Experimental Protocols

We compare the performance of the LOPA model with the Strict Binary, the Binary Majority, and the Bayesian model as well as with a ML classifier.

### *4.1. Evaluation Protocol and Majority-Class Imputation Baseline*

We test on the DIAS dataset with 30 independent, stratified random splits [24]. In each split, we hold out 20% of the cases for testing and train on the remaining 80%, preserving class proportions. This setup yields robust estimates under varying train/test partitions.

*Metrics (reported per split, then averaged).* Let $N$ be the number of test cases and let $P$ be the number of non-`NaN` predictions a model makes. We report:

- *Coverage*: $100 \times (P/N)$, the percent of test cases for which the model returns a label (non-abstention).
- *Macro–F1* [4] *(pred. only)*: macro-averaged F1 computed on the subset of test cases with non-`NaN` predictions.
- *Macro–F1 (+ $\pi$ completion)*: to compare fairly with learners that never abstain, any `NaN` prediction is replaced by a default majority-class label ($\pi$); macro–F1 is then computed on all $N$ test cases.

*Random-forest baseline.*

Following [26], which found Random Forest strongest on an alternative DIAS pruning, we use it as the sole ML baseline.

Cases are 18-bit factor vectors; hyperparameters are tuned by 5-fold CV on the training fold; the model always predicts (coverage 100%).

*Interpretation of Table 1.* Coverage varies substantially: *Strict Binary* abstains most ($\approx$69% coverage), which inflates its macro–F1 on the easier, predicted subset (0.8411), but lowers macro–F1 once abstentions are completed with $\pi$ (0.7331). After majority-

[3] Max pruning is *side-symmetric*. After deduplication, we keep *maximal* sets under $\subseteq$ on the $\pi$ (pro-plaintiff) side and *minimal* sets under $\supseteq$ on the $\delta$ (pro-defendant) side (equivalently, maximal for the $\delta$-order). Thus redundancy is removed on *both* sides—not only for $\pi$.

[4] Macro–F1 is the unweighted mean of per-class F1 scores (macro-averaging) [25]

**Table 1.** Averaged over 30 stratified 80/20 splits.

| Model | Coverage (%) | Macro–F1 (pred. only) | Macro–F1 (+ $\pi$ completion) |
|---|---|---|---|
| Bayesian Majority | 85.93 | 0.7795 | 0.7500 |
| Binary Majority | 82.93 | 0.7862 | 0.7505 |
| LOPA | 85.93 | 0.7828 | 0.7535 |
| Strict Binary | 68.70 | 0.8411 | 0.7331 |
| Random Forest | 100.00 | 0.7400 | 0.7400 |

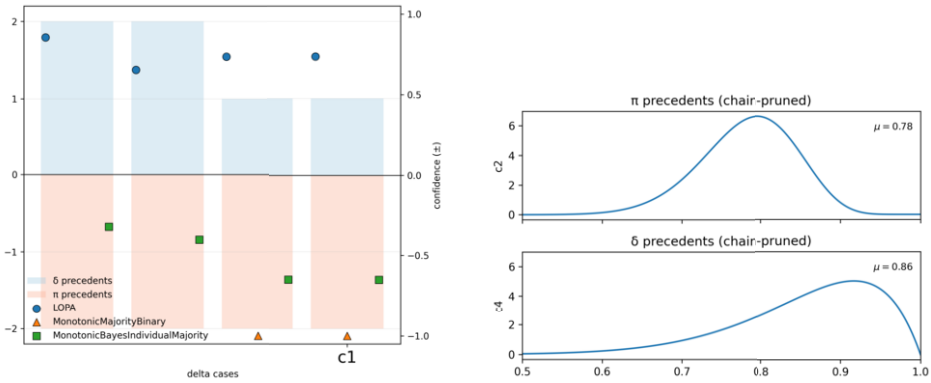



**Figure 1.** Model disagreements and an illustrative per–precedent explanation. (a) Disagreement set ($\delta$ cases). Bars (left $y$-axis) count applicable precedents—upward support $\delta$, downward $\pi$. Markers (right $y$-axis) show each model's prediction and confidence; above 0 agrees with ground truth and height encodes confidence. (b) Case $c_1$. Learned Beta posteriors for selected $\pi$ (left column) and $\delta$ (right column) precedents under LOPA.

class imputation, three symbolic models—*Naive-Bayes*, *Binary Majority*, and *Bayesian Individual Majority*—all exceed the Random Forest on macro–F1 (0.7535, 0.7505, and 0.7500 vs. 0.7400), while maintaining competitive coverage (83–86%). This suggests that the *a fortiori* structure captured by our symbolic methods aligns well with the normative regularities in the data, yielding performance on par with or better than a strong ML baseline while remaining transparent and explainable.

### *4.2. Leave–One–Out (Upper–Bound) Comparison*

After establishing parity (even slight superiority) over a strong ML baseline, we compare the symbolic models *to one another* and inspect concrete cases. To avoid dependence on random train/test splits, we use leave–one–out cross-validation (LOO) [12, §1.2.7]: for each case $i$, we train the model on all other cases $\mathscr{D} \setminus \{i\}$ and evaluate on $i$. LOO uses $n-1$ training examples per test case, and thus serves as a high-information benchmark for per-case performance (often stronger than results from smaller train/test splits).

*LOO Protocol.* Let $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^{n}$ be the dataset. For each case $j$: (i) train each model $m$ on $\mathscr{D} \setminus \{j\}$; (ii) compute the model's LOO prediction on $x_j$ (allowing abstention); (iii) for Bayesian models record also a confidence (Monte Carlo estimate).

*Where the models differ.* Figure 1(a) shows that *Bayesian Individual Majority* [9] aligns in sign with *Binary Majority* on all non-ties, but attains higher coverage by deciding exact

ties. This underscores a limitation of simple counts (Remark 1): many weak precedents can swamp a few strong ones. By contrast, our *LOPA* sometimes prefers fewer, stronger precedents and thus departs from majority when strength and count conflict. Fig. 1(b) illustrates this on the $\delta$ case `c1`: despite more $\pi$ precedents overall, the $\delta$ precedent's posterior dominates, and the model correctly predicts $\delta$.

A representative instance is the $\delta$ case labeled `c1` in Fig. 1(a), whose per-precedent posteriors are shown in Fig. 1(b).

**Example 1** (Case `c1`)**.** *Let $F(c)$ be the factor set of case $c$. Case `c1` is a $\delta$ decision with*

$$F(\texttt{c1}) = \begin{cases} f_1 = \textit{Nervous Behavior or Appearance} & f_2 = \textit{Refused Consent} \\ f_3 = \textit{Suspicious or Inconsistent Answers} & f_4 = \textit{Other} \end{cases}$$

*There are two applicable $\pi$ precedents, `c2` with $F(\texttt{c2}) = \{f_1, f_2, f_3\}$ and `c3` with $F(\texttt{c3}) = \{f_1, f_3\}$, and one applicable $\delta$ precedent, `c4` with $F(\texttt{c4}) = \{f_1, f_2, f_3, f_4\} = F(\texttt{c1})$. Since $F(\texttt{c3}) \subset F(\texttt{c2})$, max pruning discards `c3` on the $\pi$ side, keeping only the maximal subset `c2`. Fig. 1(b) shows the learned Beta posteriors for the retained precedents: the $\delta$ density (for `c4`) places most of its mass to the right of the $\pi$ density (for `c2`), reflecting greater confidence that an* a fortiori *application of `c4` supports $\delta$. Our LOPA therefore (correctly) predicts $\delta$ for `c1`. In contrast,* Binary Majority *counts two $\pi$ precedents versus one $\delta$ and would vote $\pi$, and* Bayesian Individual Majority *tends to agree with that majority sign (cf. Fig. 1(a)).*

## 5. Discussion

Beyond accuracy numbers, what makes our LOPA model distinctive is its ability to explain, justify, and calibrate its decisions. In what follows, we discuss some key aspects.

*Explainability.* ML captures correlations, not explanations; it estimates feature–label associations, multiply them, and predict. But it cannot say which precedent carried the decision or why it should be trusted. Our extended model addresses this limitation by tracing the precedents that drive the decision. In the presence of inconsistencies, each applicable precedent is treated as an explicit, auditable *a fortiori* explanation, with a learned reliability score (a Beta distribution). Decisions are then based on a small set of such explanations, with calibrated confidence. Instead of "this pattern correlates with $\pi$," we can say "this case follows precedents $\{p_1, p_2\}$ *a fortiori*, and we are $x\%$ confident."

*Compact justifications.* We prune redundant chains of precedents (e.g., when one set of facts strictly includes another). Indeed when several applicable precedents stand in an *a* fortiori chain (e.g., $S_i \subset S_j$ on the $\pi$ side or $S_i \supset S_j$ on the $\delta$ side), we keep only the extremal (maximal/minimal) elements. This yields shorter, human-scale explanations: our model cites, on average, **2.47** precedents per decision, compared to **6.14** for the Bayesian Individual Majority. This is consistent with certain judicial guidance [27, p. 159]; see also [28, p. 17], which values the weight of authorities over their number.

*Beyond means.* In the *Bayesian Individual Majority* [9], the decision depends only on the means of the per–precedent Betas. Indeed, if we write the stochastic vote margin as

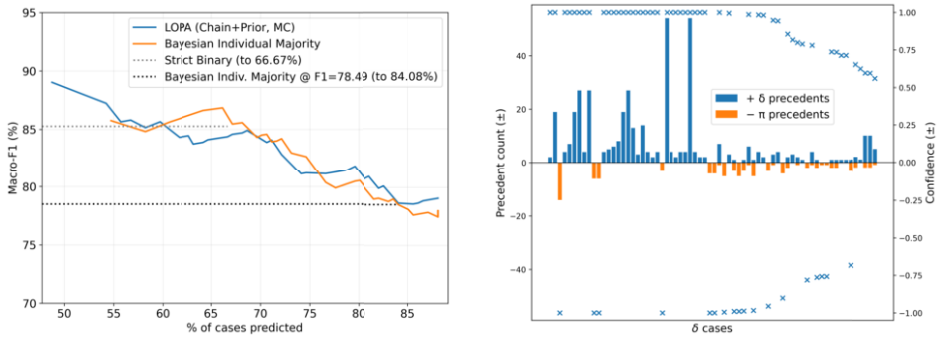



**Figure 2.** Confidence as reliability. (a) Macro–F1 versus coverage: as the confidence threshold increases, coverage drops and Macro–F1 typically rises (up to noise). (b) $\delta$ cases ordered by LOPA prediction confidence: bars show applicable precedents (upward for $\delta$, downward for $\pi$); overlaid crosses mark model predictions; Markers above 0 indicate correct predictions; those below 0 indicate errors.

$$S = \sum_{i \in E_\pi} X_i - \sum_{j \in E_\delta} Y_j, \qquad X_i \mid \Theta_i \sim \text{Bernoulli}(\Theta_i),\ Y_j \mid \Theta_j \sim \text{Bernoulli}(\Theta_j),$$

then by the linearity of expectation $\mathbb{E}[S] = \sum_{i \in E_\pi} \mu_i - \sum_{j \in E_\delta} \mu_j$ depends only on the means $\mu_i = \mathbb{E}[\Theta_i]$. *LOPA* is instead *nonlinear* in the per–precedent parameters. Writing

$$\mathscr{O}_\pi(\Theta) = \left(\frac{\pi_0}{\delta_0}\right)^{1-K} \prod_{i \in E_\pi} \frac{\Theta_i}{1-\Theta_i} \prod_{j \in E_\delta} \frac{1-\Theta_j}{\Theta_j},$$

the side and confidence are computed from

$$p_\pi = \mathbb{E}\left[\frac{\mathscr{O}_\pi(\Theta)}{1+\mathscr{O}_\pi(\Theta)}\right],$$

which *cannot* be reduced to means. Thus, while the majority model's expected margin uses means, the LOPA intrinsically leverages the full Beta distributions for both *side* and *confidence*. In either case, the reported confidence is genuinely distributional and would not be available without the Beta posteriors.

*Handling inconsistency.* A criticism to Bayesian approaches is that they are "just probabilistic." But in inconsistent corpora, all models –including Strict Binary– face uncertainty. Our Bayesian model make this uncertainty explicit, enabling principled tradeoffs between coverage and reliability.

*Confidence as a signal.* We tested whether model confidence aligns with actual reliability by sweeping a confidence threshold $\tau$ and treat predictions with confidence $< \tau$ as abstentions. As $\tau$ increases, coverage decreases, but macro–F1 improves smoothly for both Bayesian models, showing confidence is informative. Figure 2(a) shows smooth, improving macro–F1 as coverage is reduced for *both* Bayesian models, indicating that confidence is informative. Using the coverage–F1 curves, we set the smallest confidence threshold $\tau$ that matches *Strict Binary*'s $\approx 66.6\%$ coverage; ranking cases by LOPA confidence. Fig. 2(b) then shows that many one–sided cases fall *below* the threshold while

several conflicted cases remain *above*. Thus, the model abstains on potentially misleading one–sided instances yet stays confident on some conflicted ones.

In summary, our LOPA uses per-precedent distributions to determine outcome and confidence. By treating inconsistency as uncertainty, it turns confidence into an informative signal: it not only explains why a prediction is made, but also how sure the system is, providing a principled way to balance coverage against reliability.

## 6. Conclusion

We introduced a LOPA method that enhances the result model with ML techniques to effectively handle inconsistent case bases. We compared it against three existing extensions and a strong ML baseline (tuned Random Forest). All consideredsymbolic models are on par with –and in two cases slightly better than– ML when evaluated with macro–F1. In particular, our new aggregator achieved a modest performance gain w.r.t. all other models, but its main advantages are explanatory:

1. **Resistance to swamping.** Strong precedents can outweigh many weak ones, avoiding the "majority always wins" problem.
2. **Focused precedent sets.** Redundant chains of precedents are pruned, leaving concise, human-scale justifications closer to judicial practice.
3. **Quantified uncertainty.** It produces calibrated confidence scores, enabling users to abstain from low-confidence cases while maintaining high accuracy elsewhere.

We conclude that *a fortiori* precedent reasoning is competitive with ML while offering superior transparency and actionable uncertainty, properties that are essential for trustworthy legal AI. To close, we outline directions for future research, which include:

*Model generality and extensibility.* We treat the *a fortiori* constraint as one admissible rule for reasoning with precedents but the framework could also incorporate further rules, e.g. adapting argument moves from [29]. In that setting, strategies such as *downplaying a distinction* or *emphasising a strength* could be encoded as rules analogous to *a fortiori* and learned in the same way. It is also noteworthy that, while our experiments used a simplified (degenerate) instantiation of the result model, the approach is readily adapted to the full result model and, mutatis mutandis, to the reason model.

*The conditional–independence assumption* underlying Naive Bayes is only partly mitigated by max pruning; richer dependence modeling is a natural next step.

*Temporality.* The DIAS data set lacks temporal metadata, so our experiments are time-agnostic; incorporating temporality is left for future work.

*Dataset Bias.* The dataset shows two biases: one from $\pi$ being the majority class, addressed by our choice of macro–F1; another, more subtle, stems from the fact that many borderline cases are decided for $\pi$, rewarding models that lean toward the plaintiff, as gains on borderline cases outweigh the corresponding errors. In practice, the *Bayesian Individual Majority* model inherits this tilt, whereas our *LOPA* stays closer to balance. As a result, the LOPA appears only *slightly* better on DIAS, but we expect a clearer superiority on datasets without this $\pi$ bias.

**Acknowledgments:** Work supported by WWTF [Grant ID: 10.47379/ICT23030].

## References


[1] Alexander L. Constrained by Precedent'. Southern California Law Review. 1989;63:1.

[2] Horty JF. Rules and reasons in the theory of precedent. Legal theory. 2011;17(1):1-33.

[3] Horty JF, Bench-Capon TJM. A factor-based definition of precedential constraint. Artif Intell Law. 2012;20(2):181-214. Available from: https://doi.org/10.1007/s10506-012-9125-8.

[4] Horty JF. Modifying the reason model. Artif Intell Law. 2021;29(2):271-85. Available from: https://doi.org/10.1007/s10506-020-09275-z.

[5] Prakken H. A formal analysis of some factor- and precedent-based accounts of precedential constraint. Artif Intell Law. 2021;29(4):559-85. Available from: https://doi.org/10.1007/s10506-021-09284-6.

[6] Bench-Capon TJM. The Role of Intermediate Factors in Explaining Precedential Constraint. In: Grasso F, Green NL, Schneider J, Wells S, editors. Proceedings of (CMNA 2023). vol. 3614 of CEUR Workshop Proceedings. CEUR-WS.org; 2023. p. 21-32.

[7] Gray M, Savelka J, Oliver W, Ashley K. Toward Automatically Identifying Legally Relevant Factors. In: Legal Knowledge and Information Systems. IOS Press; 2022. p. 53-62.

[8] Gray M, Savelka J, Oliver W, Ashley K. Using LLMs to Discover Legal Factors. In: Legal Knowledge and Information Systems. IOS Press; 2024. p. 60-71.

[9] Morello Y, Ciabattoni A. A Bayesian View of the Result Model. Submitted. 2025. Available from: https://www.logic.at/staff/agata/submitted2025.pdf.

[10] Canavotto I. Reasoning with inconsistent precedents. Artificial Intelligence and Law. 2023:1-30.

[11] Prakken H, Ratsma R. A top-level model of case-based argumentation for explanation: formalisation and experiments. Argument & Computation. 2022;13(2):159-94.

[12] Murphy KP. Machine learning: a probabilistic perspective. MIT press; 2012.

[13] Peters JG, Bex FJ, Prakken H, et al. Justifications derived from inconsistent case bases using authoritativeness. In: Proceedings of ArgXAI 2022. vol. 3209. CEUR WS; 2022. p. 1-13.

[14] Peters JG, Bex FJ, Prakken H. Model-and data-agnostic justifications with a fortiori case-based argumentation. In: Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law; 2023. p. 207-16.

[15] Thimm M. Inconsistency measures for probabilistic logics. Artificial Intelligence. 2013;197:1-24.

[16] De Bona G, Finger M. Measuring inconsistency in probabilistic logic: rationality postulates and Dutch book interpretation. Artificial Intelligence. 2015;227:140-64.

[17] Bruninghaus S, Ashley KD. Predicting outcomes of case based legal arguments. In: Proceedings of the 9th International Conference on Artificial Intelligence and Law. ICAIL '03. Association for Computing Machinery; 2003. p. 233–242.

[18] van Woerkom W, Grossi D, Prakken H, Verheij B. A fortiori case-based reasoning: from theory to data. Journal of Artificial Intelligence Research. 2024;81:401-41.

[19] Horty JF. The result model of precedent. Legal Theory. 2004;10(1):19-31.

[20] Kaplan LM, Şensoy M, Tang Y, Chakraborty S, Bisdikian C, De Mel G. Reasoning under uncertainty: Variations of subjective logic deduction. In: Proceedings of the 16th International Conference on Information Fusion. IEEE; 2013. p. 1910-7.

[21] Kaplan L, Şensoy M, de Mel G. Trust estimation and fusion of uncertain information by exploiting consistency. In: 17th International Conference on Information Fusion (FUSION). IEEE; 2014. p. 1-8.

[22] Josang A, Ismail R. The beta reputation system. In: Proceedings of the 15th bled electronic commerce conference. vol. 5; 2002. p. 2502-11.

[23] Domingos P, Pazzani M. Beyond independence: Conditions for the optimality of the simple bayesian classi er. In: Proc. 13th Intl. Conf. Machine Learning; 1996. p. 105-12.

[24] Witten IH, Frank E. Data mining: practical machine learning tools and techniques with Java implementations. Acm Sigmod Record. 2002;31(1):76-7.

[25] Schütze H, Manning CD, Raghavan P. Introduction to information retrieval. vol. 39. Cambridge University Press Cambridge; 2008.

[26] Gray M, Savelka J, Oliver W, Ashley K. Empirical legal analysis simplified: reducing complexity through automatic identification and evaluation of legally relevant factors. Royal Society; 2024.

[27] Practitioner's Handbook for Appeals to the United States Court of Appeals for the Seventh Circuit; 2020. Revised through Sept. 28, 2020. Available from: https://www.ca7.uscourts.gov/rules-procedures/Handbook.pdf.

[28] Garner BA. Judges on Briefing: A National Survey. The Scribes Journal of Legal Writing. 2001;8. 2001–2002 issue. Available from: https://scribes.org/wp-content/uploads/2022/12/Scribes_vol8_04_Judges_on_Briefing.pdf.

[29] Bench-Capon T, Sartor G. A model of legal reasoning with cases incorporating theories and values. Artificial Intelligence. 2003;150(1-2):97-143.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*

*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251584*

# Context-Aware Citation Networks: A Human–AI Dataset, Analysis, and Tool

Jack MUMFORD [a,1], Francesco FLORIMONTE [b] Katie ATKINSON [a] and Kanstantsin DZEHTSIAROU [b]

[a] *Department of Computer Science, University of Liverpool, UK*
[b] *School of Law and Social Justice, University of Liverpool, UK*

ORCiD ID: Jack Mumford https://orcid.org/0000-0003-2467-5785, Francesco Florimonte https://orcid.org/0009-0004-1147-8496, Katie Atkinson https://orcid.org/0000-0002-5683-4106, Kanstantsin Dzehtsiarou https://orcid.org/0000-0001-9253-6109

**Abstract.** We present a context-aware approach to citation analysis for European Court of Human Rights case law. Instead of a single edge between cases, we annotate each citation instance with respect to its individual context. We construct two datasets: a new human-annotated set spanning judgments and decisions across Grand Chamber, Chamber, and Committee, and an AI-annotated set of judgments produced at scale. We find that most authorities are cited at least twice within a case, and repeated citations frequently vary by complaint and judicial consideration. Our empirical analysis shows that AI annotations are highly competitive with outputs from trained human annotators, but with significant variation depending on the provision of convention and the level of the Court producing the ruling. We release the datasets and an interactive Citation Analysis Tool that enables context-filtered retrieval, supporting triage and research grounded in past precedent.

**Keywords.** Legal judgment prediction, European Convention on Human Rights, Legal datasets, Legal citation networks

## 1. Background on AI Tools for Processing ECHR Cases

The European Convention on Human Rights (ECHR) has become a prominent domain for the development and testing of legal AI models. The Court is set up on the basis of the ECHR and adjudicates complaints concerning violations of human rights and freedoms enshrined in Articles 2-14. A key motivating factor for developing AI models for this domain is the accessibility of data on decided cases, provided through the HUDOC database.[2] The past cases stored in this repository provide source material for analysis and extraction in order to build AI models to replicate the decisions made in those cases and learn domain knowledge for deciding future cases. Prominent papers in the initial suite of this research (e.g. [1] [3] [15]) applied machine learning techniques to the task of classifying previously decided cases into the two outcome categories of violation or

[1] Corresponding Author: Jack Mumford, jack.mumford@liverpool.ac.uk
[2] https://hudoc.echr.coe.int

non-violation of an Article of the ECHR. These approaches characterise what has come to be known as *the legal judgment prediction task*. However, there are different flavours to this task and a variety of sub-tasks underpin it.

In [16] an analysis is provided of the judgment prediction task and a more nuanced distinction is made between *outcome identification* from the case texts, outcome-based *judgment categorisation* to replicate judges' decisions, and *outcome forecasting* in which the verdict is determined on the basis of textual information about the case that was available *before* the verdict was made public. Most work undertaken to date on the judgment prediction task has come under the category of judgment categorisation, aimed at demonstrating AI models' abilities to produce the same outcome as human judges did in the actual case. Whilst a significant number of contributions have used machine learning techniques to tackle the task, symbolic approaches have also been used in order to provide better explainability of the AI models' categorisation decisions [6] [17].

In order to undertake the challenging task of outcome forecasting, multiple sub-tasks need to be addressed to enable a verdict to be determined: identifying the facts of a case from claims, retrieving relevant case law, performing legal interpretation on the basis of facts and providing explanation of the conclusion reached. The work we report in this paper provides a deeper treatment of the role that citations to past case law play in contributing to a ruling on a current case.

Citations analysis for use in AI models is not a new topic of research, with prior work having been done to analyse citation networks of ECtHR case law [20], to assist with identifying landmark ECtHR cases [24] and for undertaking retrieval of prior ECtHR cases [21]. Recent legal-NLP benchmarks and multilingual resources frame our setting and future extensions ([4] [5]). For context-aware retrieval and citation modelling, we draw on case-retrieval and passage datasets and shared tasks ([9] [12] [14]) and on ECtHR-focused argument mining [10]. Foundational citation-network analyses motivate our instance-level, role-sensitive edges ([8] [13]). A recent paper [19] has focussed attention on identifying the context in which a citation appears, in order to enable the accurate extraction and contextualisation of citation instances within case law. Our work presents a new tool that has been developed for the identification of citations and, importantly, their context, within ECHR case law; whilst searching case law repositories can retrieve past cases relevant to a current case under consideration, knowing *why* a particular case was cited is important contextual detail. Coupled with the contribution of the tool itself is an annotation study that has been undertaken to manually mark up citation contexts in past cases and yield a dataset that can be used to train AI systems to take a more nuanced interpretation of the case law relevant to a new case under consideration. The ultimate aim of the tools is to provide decision support facilities for lawyers and judges processing ECHR cases.

## 2. A New Dataset for ECtHR Citation Context

This section introduces the representation and resources that underpin our study. We first define our *citation-instance* model—treating each citation as an atomic link enriched with two context dimensions, *Complaint* Article (of ECHR/Protocol) and *Judicial Consideration*—which we use consistently in later analyses (§2.1). We then describe the construction of the input corpus across court and outcome levels and the human anno-

tation workflow, including materials provided, INCEpTION-based training, and curation (§2.2). Finally, we summarise the released artefacts, outlining both the new human-annotated dataset and the AI-annotated counterpart from prior work [19] that we employ in the paper (§2.3).

### *2.1. Citation Context Modelling*

In this study, each *citation instance* is treated as an *atomic* connection between the current case document and the cited authority. This permits multiple connections between the same pair of cases within a single document whenever the authority is cited more than once. The context of each citation instance is characterised along two primary dimensions:

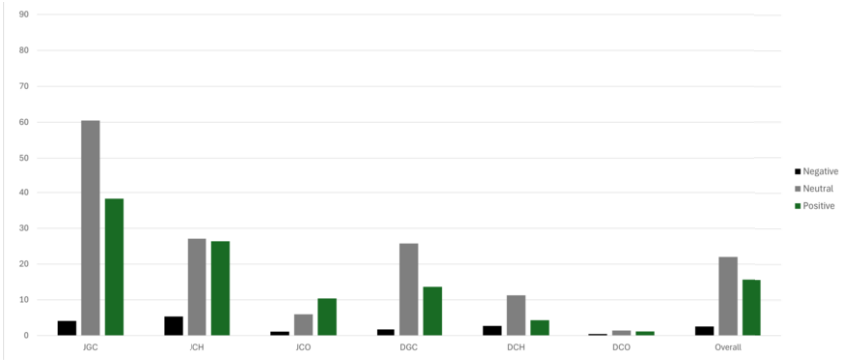
1. **Complaint (Article).** The specific Convention or Protocol provision associated with the citation of the authority.
2. **Judicial Consideration.** Whether the authority is used in a *Positive* manner (applied, approved, followed, affirmed, or otherwise adopted), a *Negative* manner (distinguished, disapproved, doubted, not followed, overruled, reversed, superseded, or otherwise rejected), or a *Neutral* manner (serving only a technical supporting role and providing no test employed by the Court to determine the outcome of the current case).

Focusing on citation *instances* (rather than document-level edges) enables a more nuanced interpretation of case law than the standard approach; it preserves the possibility that the same authority is invoked under *different complaints* and in *different judicial roles* across a single judgment. This fidelity is essential for building a database of *fine-grained connections* that captures how authorities are actually used within the Court's reasoning. Such a database supports two core use cases: **(1)**, it enables deeper exploration and analysis of the case law: by retaining multiple, context-specific links between cases, researchers can uncover patterns that are obscured by the blunt single-edge representation (e.g., when an authority is generally surveyed neutrally but later *applied* to a particular complaint). **(2)**, it provides richer supervision for AI systems: models can be trained to connect the *concrete facts* of a case to *specific citations* of authorities conditioned on their contextual usage (complaint and consideration). This, in turn, can be generalised to new cases for *context-aware authority retrieval* grounded in past precedent.

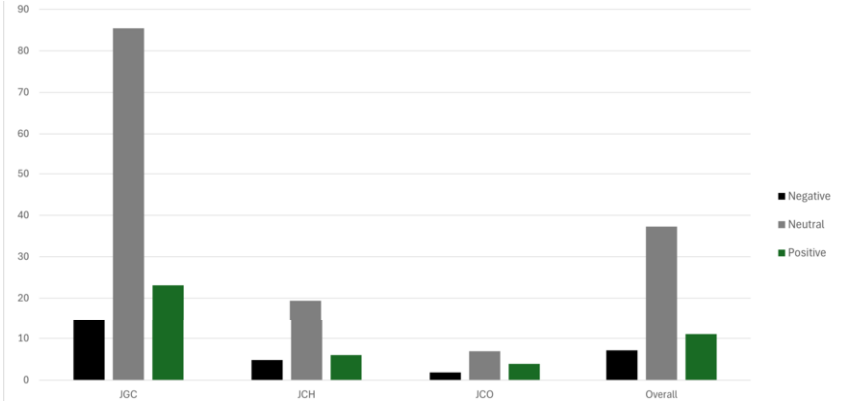
### *2.2. Input Corpus and Human Annotation Procedure*

*Corpus construction.* A rigorous selective sampling method was implemented to produce a corpus for annotation that ensured a *diversified but prevalence-aware* distribution. The corpus spans **six types of judicial output levels**—two *ruling* levels: Judgment (J), Decision (D); crossed with three *judicial formation* levels: Grand Chamber (GC), Chamber (CH), Committee (CO). Throughout the paper we denote a specific judicial output level as $[r][j]$, where $r$ is the *ruling* level, and $j$ is the *judicial formation* level.

*Annotators and materials.* Three annotators (two LLM (i.e. master of laws) students and one LLB student in Law) and one curator (a PhD Law student with substantial ECtHR research experience) conducted the human annotation. All contributors were compensated at an hourly rate for annotation and meeting time to encourage commitment




(a) Human annotations: Judgment- and Decision-level cases.




(b) AI annotations: only Judgment-level cases.

**Figure 1.** Distributions of authority-citations per case by judicial consideration label. Results are segmented by judicial output levels, where 'Overall' denotes the mean across all levels.

and quality. Annotators were provided *only* the THE LAW section of each case and instructed to annotate citations tied to the Court's *reasoning*, *omitting* citations in sections such as the parties' submissions, general principles, or other non-determinative material.

*Training, tooling, and workflow.* Training covered the INCEpTION annotation platform [11] with three practice cases pre-evaluated by the curator. The curator led the sessions, providing guidance and clarifications, and then conducted ongoing curation. After training, annotators proceeded independently, with *weekly* meetings between the team and project leadership to discuss curator-identified corrections and annotator queries. Curation aimed to ensure label consistency (particularly for *Judicial Consideration*) while preserving informative dissent when it reflected genuine ambiguity rather than error.

### 2.3. Summary of the Dataset

*Releases and composition.* We release (i) the *human-annotated* dataset produced for this study and (ii) the *AI-annotated* dataset from prior work [19] adapted for the present experiments. To preserve signal that may be smoothed away by curation, the release contains *both* the curator outputs *and* the individual annotator outputs. Disagreement can

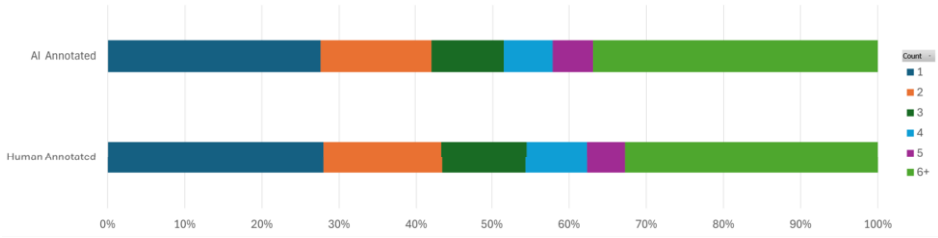



**Figure 2.** Distributions of overall counts of citations of the same authority in a given case, ranging from the proportion of authorities cited in a case that were cited only once, to those cited at least six times in the same case.

be analytically useful (e.g., identifying ambiguous passages), and retaining per-annotator labels enables future work on adjudication models and training AI from heterogeneous supervision.

*Intended uses.* Both the human-annotated[3] and AI-annotated[4] datasets support *descriptive* analysis (e.g., variance of repeated citations and consideration across judicial output levels) and *predictive* tasks (e.g., context-aware outcome aggregation and retrieval). The public artefact includes code to facilitate replication of the analyses reported in §3.

## 3. Analysing the Dataset

This section characterises the resulting citation–context dataset and evaluates how far citation context—as captured by our annotation scheme—supports downstream case-judgment classification tasks as a measure of the quality of the context captured. We report descriptive statistics (§3.1) and then compare models trained on human- versus AI-produced annotations (§3.2).

### *3.1. Annotation Data Statistics*

*Data distributions.* Human annotators covered *judgments* and *decisions* at GC (27 judgments, 8 decisions), Chamber (26 judgments, 7 decisions), and Committee (40 judgments, 7 decisions) levels, for a total of 115 cases. The AI annotations were generated only for judgments (100 per level; 300 in total), as reported in previous work [19], to facilitate outcome classification (violation or nonviolation) as a proxy task to indicate citation context quality. We can therefore assess citation–context quality indirectly by comparing outcome accuracy.[5] Human annotation likewise prioritised judgments for the same reason, but *decision* cases (admissibility) were also included to broaden dataset coverage.

*Judicial consideration by judicial output level.* Figure 1 (left: human, 1a; right: AI, 1b) shows very similar distributions overall, with *Neutral* the most prevalent label across judicial output levels. Two systematic differences emerge from the human annotations:

[3]Human-annotated: https://github.com/jamumford/ECHR-citation-context-v1
[4]AI-annotated: https://github.com/jamumford/ECHR_Citation_Context
[5]We return to the predictive evaluation in §3.2.

(i) *Positive* consideration is ascribed more frequently than in the AI outputs, and (ii) for *Judgment–Committee* cases, *Positive* even *surpasses Neutral*. Consistent with expectations, total citation counts increase with higher judicial output levels. From the human chart we also see that citations per case are higher for *Judgment* cases than for *Decision* cases.

**Table 1.** Percentage of authorities cited with different contexts within the same case: judicial consideration; and associated complaint (Article of the ECHR/Protocol). Results compare human- and AI-annotation across different judicial output levels. Accuracy reported to 2 sf.

| | **Judicial Consideration** | | **Associated Complaint** | |
|---|---|---|---|---|
| **Judicial Output Level** | **Human %** | **AI %** | **Human %** | **AI %** |
| Judgment Grand Chamber | 66 | 71 | 54 | 46 |
| Judgment Chamber | 67 | 48 | 41 | 41 |
| Judgment Committee | 52 | 44 | 56 | 41 |
| Decision Grand Chamber | 51 | – | 37 | – |
| Decision Chamber | 56 | – | 22 | – |
| Decision Committee | 100 | – | 0 | – |
| Overall | 63 | 64 | 48 | 45 |

*Evidence of the merit of the approach.* **First**, the citation–counts chart (Figure 2) show that a clear majority of cited authorities are referenced *more than once* within the same case, and *almost one third* are cited at least six times. This pattern is consistent in both human- and AI-annotated outputs. **Secondly**, Table 1 indicates that, among authorities cited repeatedly, *most* (roughly two thirds) are accompanied by *different judicial–consideration* labels across mentions, which is exactly the kind of contextual variability our scheme is designed to capture. **Thirdly**, the same table shows that a substantial (though not quite majority) proportion of repeat citations are associated with *different complaints* (i.e., different Convention provisions), further underscoring the value of modelling citation context rather than treating citations as homogeneous units.

**Table 2.** Human annotator agreement across the different labelling categories indicating the mean Krippendorff's Alpha ($\alpha$) annotator agreement with the curator. Label categories: **P** = Position, **AC** = Associated Complaint, **CN** = Case Name, **AT** = Authority Type, **JC** = Judicial Consideration. Accuracy reported to 2 sf. Best results for each judicial output level row in **bold**.

| **Judicial Output Level** | **P** | **AC** | **CN** | **AT** | **JC** |
|---|---|---|---|---|---|
| Grand Chamber | 94 | 73 | 85 | 94 | 69 |
| Chamber | **97** | **80** | **87** | **96** | **70** |
| Committee | 94 | 66 | 81 | 94 | 59 |

*Annotator agreement.* Table 2 reports agreement by judicial output level. Somewhat counter-intuitively, the *Chamber* level yields the highest agreement—suggesting a "Goldilocks" zone between the comparatively sparse and inconsistently descriptive *Committee* material and the richly detailed but *complex Grand Chamber* decisions. As expected, categories that require more substantive reasoning show lower agreement than simpler descriptive labels. Nevertheless, agreement is *high across the board*; given that

the agreement metric accounts for label distributions, these scores are strong even under a conservative calculation.

### 3.2. Human vs AI Annotation Quality

*Method recap.* We apply the method used in [19]: using judgment classification experiments as a proxy for evaluating the accuracy of inferred citation contexts. The intuition is that: successfully capturing the context of the Court's citation of authorities $\Longrightarrow$ accurate case judgment classification, which means that we can use any improvement in case judgment classification accuracy to infer backwards that we have more accurately captured citation context. We predict case complaint outcomes by aggregating over the outcomes of the *cited authorities*. For each complaint Article in the current case, we keep running counts of VIOLATION vs. NON-VIOLATION among the cited authorities and take the majority as the predicted outcome. Crucially, *citation context* enables a principled filter: counts are incremented *only when* the cited authority is applied to the same complaint with *Positive* judicial consideration. The *baseline* classifier offers no such filtering—reflecting traditional citation-frequency approaches that ignore (i) whether the citation is supportive and (ii) whether it applies to the same complaint.

**Table 3.** Judgment classification F1 scores for specific complaint outcomes (VIOLATION or NONVIOLATION). Comparison of classification based on human- vs AI-annotated cases for both their respective full datasets, and the 27 intersecting cases dataset. Benchmark classifier from [19] is reproduced for human-annotated full dataset and intersecting dataset. Results normalised to range 0-100. Accuracy reported to 1 dp. Best results for each judicial output level row in **bold**.

| | Full Datasets | | | Intersecting Dataset | | |
|---|---|---|---|---|---|---|
| | **Human F1** | **Baseline F1** | **AI F1** | **Human F1** | **Baseline F1** | **AI F1** |
| JGC | 75.4 | 64.4 | **76.5** | – | – | – |
| JCH | 69.9 | 76.0 | **83.6** | – | – | – |
| JCO | 45.5 | 47.3 | **66.6** | – | – | – |
| Overall | 69.0 | 68.2 | **76.5** | **74.6** | 64.1 | 71.1 |

*Full datasets: human vs. AI vs. baseline.* Table 3 reports F1 for models trained on (a) human-annotated contexts, (b) the best-performing AI-annotated contexts, and (c) the baseline. Across the full datasets, the *AI* setting yields the strongest performance overall, with the baseline (lacking context filtering) coming bottom.

*Intersecting dataset (27 cases).* We also evaluate on the 27 judgments annotated by both sources (23 JGC, 1 JCH, 3 JCO). On this *intersecting* set, the *human* annotations achieve the best overall performance, with human vs. AI very close at JGC level, while *AI clearly exceeds human* at JCH and JCO levels. The human win on the intersecting set may be driven by its strong weighting toward JGC cases. Consistent with [19], the baseline is *overall* outperformed by both AI- and human-annotation-derived context.

*By Article.* Table 4 breaks down F1 by Convention/Protocol provision. Some complaints (most notably *Article 5*) yield similar returns across human- and AI-annotated datasets, but most provisions show *substantial* divergence between the two. *Articles 3 and 8* are the only provisions where human-annotated outputs produce superior classification performance; AI leads elsewhere.

**Table 4.** Judgment classification F1 scores for specific complaint outcomes (Articles/Protocols violated or not violated) for respective human- and AI-annotated datasets. Accuracy reported to 1 dp. Best results for each Article column in **bold**.

| | 2 | 3 | 5 | 6 | 8 | 10 | 11 | 13 | 14 | P1-1 |
|---|---|---|---|---|---|---|---|---|---|---|
| Human | 50.0 | **66.7** | 56.4 | 49.7 | **82.9** | 43.7 | 49.7 | 56.8 | 55.6 | 32.0 |
| AI | **61.4** | 60.7 | **56.6** | **76.2** | 70.0 | **79.0** | **57.5** | **66.7** | **63.3** | **74.6** |

## 4. A Tool for Citation Context Analysis

*Overview.* The tool incorporates datasets built from annotated case law, allowing users to engage in exploratory work (e.g., doctrinal mapping) and targeted retrieval (e.g., locating authorities positively applied to a given complaint).

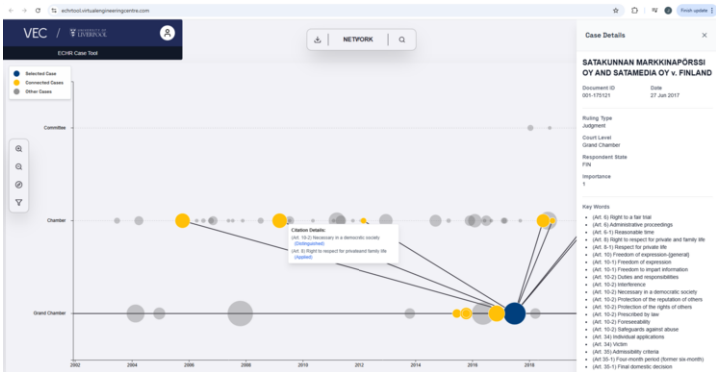



**Figure 3.** Citation Analysis Tool incorporating rich citation context, including across repeated citations of the same authority by the same case.

*Context-aware retrieval.* Unlike traditional citation search, the tool enables users to filter searches to those cases *cited within a specific context*. Filtering is not limited to the complaint linked to the citation; it also captures *how* the citation functions in the Court's reasoning via the *Judicial Consideration* label. Concretely, users can: (i) constrain by judicial output level and decision type, (ii) select the relevant Article complaint, (iii) select only those cases corresponding to desired importance levels designated by the Court, and (iv) restrict results to citations with *Positive* consideration to surface supportive applications, or include other considerations (such as *Negative*) to study contrasting uses.

*Workflow and outputs.* Users can download a table in csv format of any subgraph of citation connections from any focal case, showing the citing case, the complaints, the consideration labels, and any other selected features (such as importance or complaint outcomes). This design enables common research tasks: identifying authorities repeatedly and positively applied to a complaint; comparing treatment across judicial output levels; or tracing where the same authority is applied differently across complaints.

## 5. Discussion and Conclusions

*Evaluating the dataset.* Across both annotation sources, we find that the *majority* of authorities cited in a case are cited at least twice, and a substantial proportion of these repeatedly cited authorities are invoked under *different* Article complaints and with *different* judicial considerations (Table 1). This provides strong evidence for modelling *citation context*: if a network uses a single undifferentiated edge between two cases, it collapses multiple, distinct argumentative roles into one and discards signal that is predictive of outcomes and reasoning structure. The human- and AI-derived datasets display *similar distributions* (§3), supporting the robustness of these findings about the spread of context and court-level differences. We also observe a consistent court-level effect: the "higher" the court, the *more* authorities are cited and the *greater* the share labelled *Neutral*. This is intuitive—Grand Chamber and Chamber matters tend to be more complex/important, prompting broader canvassing of authority even where an authority is not ultimately applied or distinguished.

*AI vs. Human.* The continuing effectiveness of human-led annotation has been called into question (e.g., [18] [22] [23]). Using OpenAI GPT-4o–family models, annotating *300* judgments with AI cost roughly *2%* of the spend required for the human-annotated set of *115* cases. We are also trialling voluntary human annotation to reduce direct financial costs to zero (rendering time the sole tangible cost). On performance, AI surpassed human on the *full* datasets with a very small gap at Grand Chamber level and AI conspicuously stronger at Chamber and Committee, while human annotations led on the *intersecting* 27-case set (Table 3; see also the provision-level view in Table 4). We will test more recent and open-source models; note, however, that in our setting GPT-4o-mini outperformed GPT-4o, so gains from "larger/newer" models are not guaranteed.

*AI* with *Human.* Provision-level results (Table 4) show where humans excel and where AI does—AI wins overall on the full set, but human wins on the like-for-like intersecting set. A natural path is to *train* (and evaluate) AI models on the human-annotated data. An alternative that may scale better is *AI-first, human-refine*: generate initial AI annotations, use humans to correct/confirm difficult spans and edge cases, and iterate. This bootstrapping could stabilise provision-specific weaknesses while preserving AI's coverage advantages.

*Citation Analysis Tool and AI: next steps.* Our evaluation to date used majority-based prediction of complaint outcomes conditioned on context (§3.2). With evidence that contextualisation tracks the Court's reasoning, we now move toward *retrieval*: training and testing models to surface authorities *relevant to a given complaint and consideration*, before applying them to the *facts* of a new case to support triaging and legal research. We will also explore richer contextualisation: (i) finer-grained judicial-consideration beyond three labels, including *magnitude/position* signals (e.g., end-of-analysis citations may be more dispositive); (ii) integration of *case-importance* signals [2] (e.g., external importance labels or graph-based centrality already visualised as node size in the tool); and (iii) connection to structured argument/decision models for specific Articles/Protocols to yield suggested lines of reasoning. Ultimately, we want the tool to suggest candidate authorities *from features alone*, without first loading a source case, and to accept *fact patterns* as input to produce relevant case law and a sketched reasoning path (starting with Article 3 as a proof of concept).

*Limitations.* (i) The AI component currently uses only GPT-4o–based models; we will evaluate newer and open-source alternatives. (ii) There is no true "gold" ground truth for citation context—even judicial or practitioner annotations would reflect subjective judgement; the need for caution in treating judicial opinions as data for training legal AI is discussed in [7]—hence our use of judgment classification as a pragmatic proxy (See Method recap from §3.2). (iii) Our analysis of human-annotation is based on the curated outputs, which cannot be guaranteed to offer perfect resolution of disagreement. For this reason, we ensure our datasets also contain the raw annotations, enabling further investigation. (iv) We process only *English*-language cases, which nevertheless provide substantial ECtHR coverage. (v) Complaint context is modelled at the *Article* level; the Court often reasons at finer granularity (paragraphs/sections). We avoided this to keep human annotation tractable, but AI performance at this finer level is a promising extension.

*Conclusion.* We introduced a *context-aware* citation-analysis prototype and its underpinning dataset for navigating ECHR case law. We showed that repeated citation is prevalent (approximately two thirds of cited authorities are cited at least twice – Figure 2) and that repeat citations frequently vary by context (roughly half differ by Article and around two thirds differ by judicial consideration – Table 1). Traditional one-edge citation networks miss this structure; our approach retains it and promises to be empirically useful for outcome prediction and retrieval. The tool is operational and will continue to be developed to support efficient triaging and deeper, context-sensitive legal research at the ECtHR.

## Acknowledgements

The work reported in this paper was supported by funding from the UK's EPSRC and AHRC through their Impact Acceleration Award scheme. We also acknowledge the development of the citation analysis tool by the Virtual Engineering Centre (VEC) team in the Digital Innovation Facility (DIF) at the University of Liverpool. We are grateful to Frankie Dakin for her contributions to the analysis of data presented in this paper.

## References

[1] N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, and V. Lampos. Predicting judicial decisions of the ECHR: A natural language processing perspective. *PeerJ Computer Science*, 2:e93, 2016.

[2] D. Bareham, K. Atkinson, J. Mumford, and J. Marshall. Curb your enthusiasm: Towards a rag framework to forecast case importance in the echr. In *Proceedings of the 20th International Conference on Artificial Intelligence and Law, ICAIL 2025, Chicago, USA*. ACM Press, 2025.

[3] I. Chalkidis, I. Androutsopoulos, and N. Aletras. Neural legal judgment prediction in english. *arXiv preprint arXiv:1906.02059*, 2019.

[4] I. Chalkidis, M. Fergadiotis, and I. Androutsopoulos. Multieurlex-a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6974–6996, 2021.

[5] I. Chalkidis, A. Jana, D. Hartung, M. Bommarito, I. Androutsopoulos, D. Katz, and N. Aletras. Lexglue: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, 2022.

[6] J. Collenette, K. Atkinson, and T. Bench-Capon. An explainable approach to deducing outcomes in European Court of Human Rights cases using ADFs. In *Proc. of COMMA 2020*, pages 21–32, 2020.

[7] C. Cox. Non-herculean data. a philosophical intervention in a technical debate about judicial opinions as data sources. In *Proceedings of the 20th International Conference on Artificial Intelligence and Law, ICAIL 2025, Chicago, USA*. ACM Press, 2025.

[8] J. H. Fowler, T. R. Johnson, J. F. Spriggs II, S. Jeon, and P. J. Wahlbeck. Network analysis and the law: Measuring the legal importance of precedents at the us supreme court. *Political Analysis*, 15(3):324–346, 2007.

[9] R. Goebel, Y. Kano, M.-Y. Kim, J. Rabelo, K. Satoh, and M. Yoshioka. Overview of benchmark datasets and methods for the legal information extraction/entailment competition (coliee) 2024. In *JSAI International Symposium on Artificial Intelligence*, pages 109–124. Springer, 2024.

[10] I. Habernal, D. Faber, N. Recchia, S. Bretthauer, I. Gurevych, I. S. genannt Döhmann, and C. Burchard. Mining legal arguments in court decisions. *Artificial Intelligence and Law*, 32(3):1–38, 2024.

[11] J.-C. Klie, M. Bugert, B. Boullosa, R. E. de Castilho, and I. Gurevych. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics, June 2018. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

[12] H. Li, Y. Shao, Y. Wu, Q. Ai, Y. Ma, and Y. Liu. Lecardv2: A large-scale chinese legal case retrieval dataset. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2251–2260, 2024.

[13] Y. Lupu and E. Voeten. Precedent in international courts: a network analysis of case citations by the european court of human rights. *British Journal of Political Science*, 42(2):413–439, 2012.

[14] Y. Ma, Y. Shao, Y. Wu, Y. Liu, R. Zhang, M. Zhang, and S. Ma. Lecard: a legal case retrieval dataset for chinese law system. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2342–2348, 2021.

[15] M. Medvedeva, M. Vols, and M. Wieling. Using machine learning to predict decisions of the European Court of Human Rights. *Artificial Intelligence and Law*, pages 1–30, 2019.

[16] M. Medvedeva, M. Wieling, and M. Vols. Rethinking the field of automatic prediction of court decisions. *Artificial Intelligence and Law*, 31(1):195–212, 2023.

[17] J. Mumford, K. Atkinson, and T. Bench-Capon. Combining a legal knowledge model with machine learning for reasoning with legal cases. In M. Grabmair, F. Andrade, and P. Novais, editors, *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, ICAIL 2023, Braga, Portugal, June 19-23, 2023*, pages 167–176. ACM, 2023.

[18] J. Mumford, K. Atkinson, and T. Bench-Capon. Human performance on the ai legal case verdict classification task. In *Proceedings of JURIX 2023*, 2023.

[19] J. Mumford, D. Bareham, K. Atkinson, and J. Marshall. Finding the goldilocks zone: Retrieving citation context. In *Proceedings of the 20th International Conference on Artificial Intelligence and Law, ICAIL 2025, Chicago, USA*. ACM Press, 2025.

[20] H. Olsen and A. Küçüksu. Finding hidden patterns in ecthr's case law: On how citation network analysis can improve our knowledge of ecthr's article 14 practice. *International Journal of Discrimination and the Law*, 17:4–22, 2017.

[21] T. Y. S. S. Santosh, R. Haddad, and M. Grabmair. Ecthr-pcr: A dataset for precedent understanding and prior case retrieval in the european court of human rights. In *Proc. of Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 5473–5483. ELRA and ICCL, 2024.

[22] J. Savelka. Unlocking practical applications in legal domain: Evaluation of GPT for zero-shot semantic annotation of legal texts. In M. Grabmair, F. Andrade, and P. Novais, editors, *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, ICAIL 2023, Braga, Portugal, June 19-23, 2023*, pages 447–451. ACM, 2023.

[23] J. Savelka and K. D. Ashley. The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts. *Frontiers in Artificial Intelligence*, 6:1279794, 2023.

[24] G. van Dijck, B. Rodrigues de Miranda, and C. Crombach. Centrality scores and precedent value in legal network analysis. In *Legal Knowledge and Information Systems - JURIX 2023*, pages 247–256. IOS Press, 2023.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251585*

# PersonaMatrix: A Recipe for Persona-Aware Evaluation of Legal Summarization

Tsz Fung PANG[a,1], Maryam BERIJANIAN[b], Thomas ORTH[a], Breanna SHI[a], Charlotte S. ALEXANDER[a]
[a] *Georgia Institute of Technology*
[b] *Michigan State University*

**Abstract.**
Legal documents are often long, dense, and difficult to comprehend, not only for laypeople but also for legal experts. While automated document summarization has great potential to improve access to legal knowledge, prevailing task-based evaluators overlook divergent user and stakeholder needs. Tool development is needed to encompass the technicality of a case summary for a litigator yet be accessible for a self-help public researching for their lawsuit. We introduce *PersonaMatrix*, a persona-by-criterion evaluation framework that scores summaries through the lens of six personas, including legal and non-legal users. We also introduce a *controlled dimension-shifted* pilot dataset of U.S. civil rights case summaries that varies along depth, accessibility, and procedural detail as well as *Diversity-Coverage Index (DCI)* to expose divergent optima of legal summary between persona-aware and persona-agnostic judges. This work enables refinement of legal AI summarization systems for both expert and non-expert users, with the potential to increase access to legal knowledge. The code base and data are publicly available in GitHub[2].

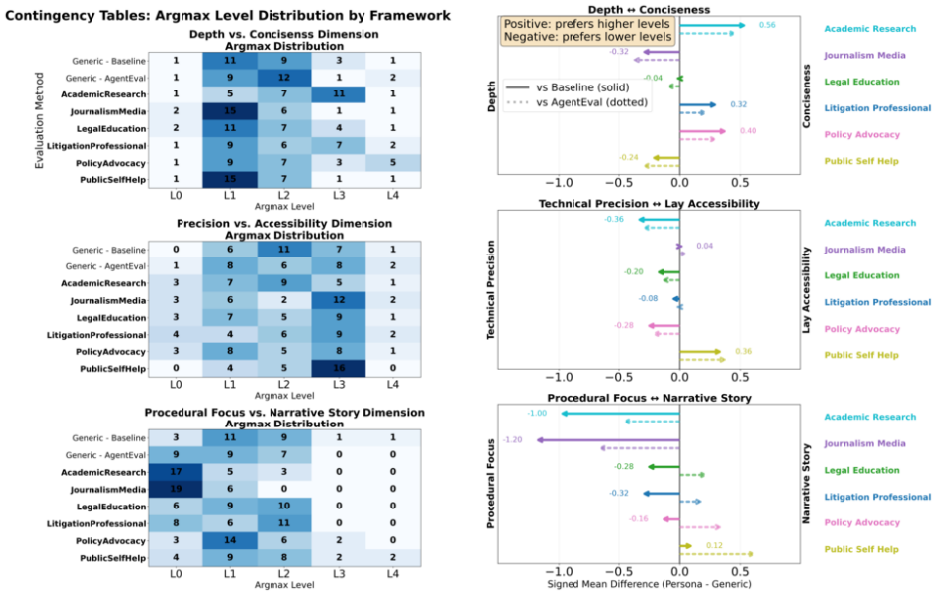
**Keywords.** legal summarization, evaluation metrics, large language models, persona-based evaluation, human-centered AI, stakeholder utility.

## 1. Introduction

Automatic summarization of lengthy, complex legal documents can expand access and comprehension for diverse end users. Although large language models (LLMs) have advanced fluency and coherence, evaluating whether a summary meets user needs remains challenging. Traditional metrics such as ROUGE [1], BLEU [2], METEOR [3], and BERTScore [4] measure lexical or semantic similarity to references, but such overlap is an unreliable proxy for utility. N-gram methods can only reward wording mimicry and even favor incoherent outputs [5,6]; embedding-based methods like BERTScore improve semantic matching yet still prioritize closeness to a reference over *fitness for task*.

As LLMs are deployed in domain-specific settings, LLM-as-a-judge and agentic-judge frameworks have emerged as alternatives. Structured designs such as AgentEval

[1] Corresponding Author: Tsz Fung Pang, tpang34@gatech.edu
[2] https://github.com/sampangtf/persona-matrix




(a) Optimal Shift Level Heatmap (b) Persona Judges' Deviated Optima from Generic

**Figure 1.** (a) Distribution of optimal (argmax) levels along three summary quality dimensions (Depth vs. Conciseness, Technical Precision vs. Lay Accessibility, and Procedural Focus vs. Narrative Story) for 6 persona judges and 2 generic judges over 25 civil rights cases. (b) Direction and magnitude of deviation of persona-aware judges' mean optimal levels from generic judges.

[7] stabilize judgments and improve explainability, sometimes correlating better with ground truths. Yet these methods assume a single notion of utility. In legal summarization, stakeholders have divergent priorities, and collapsing judgments into a scalar score with standardized criteria obscures real trade-offs. We therefore introduce ***PersonaMatrix***, a persona-by-criterion evaluation framework that fills the gap by capturing multi-objective utility and distinct priorities across heterogeneous stakeholders. It operationalizes persona-centered utility via scalable persona-conditioned evaluators, offering a middle ground between costly user studies and generic automated metrics.

Stakeholder-sensitive evaluation requires datasets where the same case is summarized along controlled quality dimensions. In this paper, a *quality dimension* refers to a specific controllable aspect of writing that characterizes a summarization, including but not limited to lexical choice, stylistic register, compression ratio and content selection. To probe trade-offs systematically, we construct a ***controlled dimension-shifted dataset*** of U.S. civil rights case summaries, where each case is summarized along conflicting quality dimensions. A scalable LLM-driven pipeline generates and validates these controlled variants from human-written seed summaries, as shown in Figure 2, enabling sensitivity analysis at manageable cost. We instantiate six personas as representative stakeholders: (1) Litigator, (2) Legal Educator, (3) Journalist, (4) Self-Help Public, (5) Academic Researcher, and (6) Policy Advocate. For example, litigators require procedural precision and citations, whereas journalists or public-facing communicators prioritize clarity, brevity, and plain language. Documentation for other persona designs, the full code base and data are available on GitHub. Gains for one group often entail losses for an-

other: procedural detail helps lawyers but burdens general readers; simplifying language aids readability but can obscure legal reasoning, controlling rules or citations. Figure 1 shows our evaluation results over the pilot datasets, demonstrating deviated optima of persona-aware judges relative to generic judges: baseline (vanilla LLM-as-a-judge) and AgentEval.

We further propose the ***Diversity-Coverage Index (DCI)***, a bounded metric that measures how well an evaluation framework preserves between-persona diversity while maintaining within-persona coherence. DCI benchmarks persona-aware evaluation and exposes limitations of single-perspective judges. We show that persona-conditioned evaluation diverges in a statistically significant way from persona-agnostic judges, revealing systematic biases and stakeholder-specific utility gaps, and enabling multi-lens reporting and iterative refinement to make legal summaries more useful for a wide range of users with differing informational needs and levels of legal expertise, thereby supporting equal accessibility of legal knowledge.

*Contributions.* The key contributions are:

1. Formalize legal-summary evaluation as a multi-objective problem across distinct stakeholders.
2. Assemble a controlled dimension-shifted dataset of U.S. civil rights case summaries that systematically varies summaries along three conflicting quality dimensions to probe stakeholder trade-offs, namely depth $\leftrightarrow$ conciseness, technical precision $\leftrightarrow$ lay accessibility, and procedural detail $\leftrightarrow$ narrative story.
3. Present PersonaMatrix, a persona-aware evaluation framework that couples persona-conditioned judges with the dimension-shifted dataset.
4. Introduce a novel metric, Diversity-Coverage Index (DCI), to quantify an evaluation framework's ability to capture between–persona distinctness and within-persona coherence.
5. Provide empirical comparison between persona-aware and persona-agnostic evaluators, demonstrating statistically significant divergences in stakeholder needs and revealing systematic biases in generic evaluators.

## 2. Related Work

Recent AI&Law work demonstrates the importance of persona-driven, user-centered design for legal AI systems [8,9], supported by findings of wide gaps in usability difficulties between non-lawyers and lawyers [10]. Bex (2024) explicitly calls for evaluating AI legal systems with stakeholders from practice [11]. Hagan (2024) advocates for multi-stakeholder criteria-based evaluation for Legal Q&A systems with the support of human user studies [12]. While recent legal summarization systems target plain-language, accessible summaries for non-lawyers [13], we still lack evaluation frameworks that assess fitness across stakeholder groups. We fill these gaps by operationalizing persona-driven evaluation for legal summarization, enabling automatic feedback for continual refinement of legal summarization systems for all stakeholders.

LLM-as-a-judge evaluates summaries with rubric-guided scoring [14], but is sensitive to prompt design and exhibits well-documented positional and verbosity biases [15]. Agentic frameworks such as AgentEval decompose evaluation into criteria genera-

tion and scoring, improving interpretability [7]. However, both assume a single optimal rubric, ignoring divergent stakeholder needs in domains like legal summarization. PersonaMatrix addresses these shortcomings by accounting for distinct informational priorities across personas and expanding evaluation beyond persona-agnostic biases.

A parallel thread in NLP and Human-Computer Interaction (HCI) has shifted system evaluation toward assessing whether outputs support decision making and task completion, aligning with usability and user experience principles [16,17,18]. Although user studies remain the "gold standard", they are costly and slow, limiting scalability.

In HCI, *personas* serve as archetypes grounding design in concrete user needs [19,20]. In NLP, personas have primarily shaped dialogue generation [21], with little attention to their role as *evaluation lenses*. PersonaMatrix is intended to uncover persona-aware utility.

Analyzing stakeholder-sensitive evaluation requires datasets where the same case is summarized along controlled quality dimensions. Existing legal datasets such as MultiLexSum [22] and BillSum [23] are valuable but lack systematic dimension-shifting for summaries of the same case. MultiLexSum varies summary length, but does not capture trade-offs beyond granularity. Outside law, PersonalSum [24] offers personalized news summaries conditioned on reader profiles, but this targets individual preferences rather than conflicting, role-based notions of utility. To fill this gap, we introduce a *controlled dimension-shifted* dataset of U.S. civil rights cases, explicitly varying summaries along conflicting quality dimensions to expose cross-persona trade-offs.

## 3. Methodology

### *3.1. Controlled Dimension-Shifted Dataset*

We design a scalable LLM-driven generator that transforms a seed summary into variants shifted along stakeholder-prioritized dimensions. This produces sets of summaries of the same case with systematically varied quality dimensions, avoiding bias from differences in case complexity. These controlled variants expose persona-specific preferences along dimension-shift spectra, revealing distinct and diverse optima. We test two hypotheses:

**Hypothesis 1 (H1)** *PersonaMatrix produces distinct, illustrative evaluation results across personas, revealing multiple optima that differ in directions and magnitudes. In contrast, persona-agnostic frameworks (LLM-as-a-judge, AgentEval) collapse to a scalar score from a single perspective, as shown by the Diversity-Coverage Index (DCI).*

**Hypothesis 2 (H2)** *The diversity detected by persona-aware evaluation reflects consistent persona signals rather than random noise under corruption tests.*

We identified a list of dimensions of legal summary qualities that are in conflict with each other and are prioritized by distinct stakeholders: (1) Depth vs. Conciseness, (2) Technical Precision vs. Lay Accessibility, and (3) Procedural Focus vs. Narrative Story. We chose these dimensions because they can be defined objectively and validated with measurable metrics, enabling automatic and scalable transformation and validation, and they are likely differentially prioritized by different stakeholders, manifesting the dilemma in optimizing the summary for multi-stakeholders, as shown in the rightmost

**Table 1.** Three quality dimensions and levels. Levels are ordinal and monotone along each axis, higher levels move further toward the right-hand label. Level 0 is the expert-authored original summary; Levels 1-4 are validated rewrites from our generator. Arrows denote the direction of shift. The last row shows how each dimension shift is validated across levels.

| Level | Depth ⇒ Conciseness | Technical Precision ⇒ Lay Accessibility | Procedural Focus ⇒ Narrative Story | Typical Stakeholder Alignment† |
|---|---|---|---|---|
| 0 | >500 words; original expert summary | Legal jargon | Full procedure | Legal professionals |
| 1 | Condense sentences; prune redundancies | Professional summary | Explain rationale | Academics |
| 2 | Summarize procedural cluster | Plain language with explanation | Only central legal issues | (Varies) |
| 3 | Executive brief (skim-first) | Simple analogies | Outcome-first; minimal procedure | Non-legal users |
| 4 | Headlines & bullets; very short | Basic vocabulary; no jargon | Narrative human story | Public |
| **Val metric** | Word count falls in a stricter band than the previous level[a] | Readability grade strictly decreases vs. previous level[b] | Count of procedural markers decreases vs. previous level[c] | N/A |

† Alignment is illustrative, not prescriptive.
[a] We enforce monotone concision by pre-specified length bands (Level $k$ must be strictly shorter than Level $k-1$).
[b] Readability is computed with standard formulas. A lower grade indicates simpler prose; Level $k$ must be strictly lower than Level $k-1$.
[c] Procedural markers include docket/procedure cues. We count unique occurrences from a fixed list; Level $k$ must score lower than Level $k-1$.

column in Table 1. The typical stakeholder alignment is intended as *refutable intuition*, not fixed mapping.

We developed a scalable dataset generator using an `Extractor → Rewriter → Validator` agentic pipeline (Figure 2-top) that shifts summaries along the three quality dimensions while enforcing factual consistency and progressive change through a multi-step validation process, including metric-based checks and LLM-powered pairwise critiques. Rewriting prompts and validation logic are detailed on GitHub.

Seed summaries are drawn from the Civil Rights Litigation Clearinghouse[3], authored and validated by legal experts–law professors and law students at the University of Michigan Law School. These texts emphasize depth, technical precision, and procedural details, enabling unidirectional quality shifts along our selected dimensions. We provide an initial probe dataset of 25 cases ($\times$ 3 dimensions $\times$ 4 levels = 300 shifted summaries with 25 originals) together with code for reproducibility.

### 3.2. PersonaMatrix Evaluator

We adapt AgentEval's [7] Critic-Quantifier pipeline to a persona-aware setting. The ***Critic agent*** defines task-based criteria and acceptable values; the ***Quantifier agent*** scores inputs against them (Figure 2-bottom).

Let $\mathscr{P} = \{\text{Litigators, Public, Journalists, Researchers, Educators, Advocates}\}$ denote persona judges. For each persona $p \in \mathscr{P}$, a Persona-Critic generates a rubric conditioned on persona-specific context for $p$. Each criterion $c$ defines accepted values (ordinal labels with descriptions). Full implementation can be found in the GitHub repo.

Rubrics are cached per persona and reused across cases. On top of the task-based evaluation criteria generation prompt suggested in AgentEval, we append the persona name and one-sentence intended usage to the Critic agent's system prompt. We demonstrate that minimal context suffices to elicit stakeholder-specific criteria. Then, a Quantifier agent assigns both ordinal labels and numeric scores $s_{p,c} \in [0,5]$ for each criterion. Scores are aligned with ordinal classes (e.g., top class $\approx$ 4.5–5), requiring no rescaling.

[3] https://clearinghouse.net/

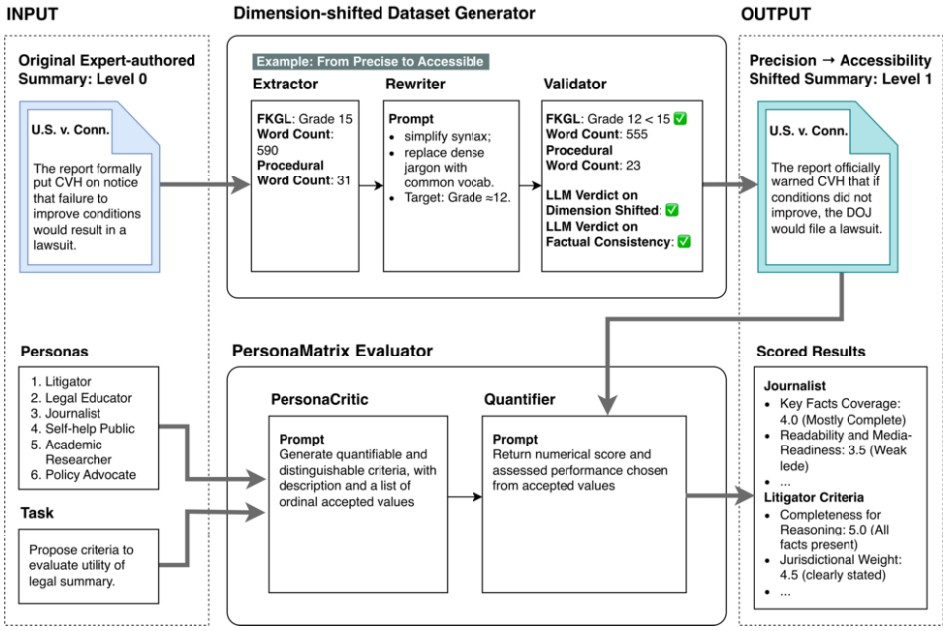



**Figure 2.** Demonstrations of how legal summary is rewritten in Controlled Dimension-Shifted Dataset Generator using the shift from Precision to Accessibility as example (top), and scored in PersonaMatrix Agentic Evaluator (bottom). There are $n$ PersonaCritics generating $n$ sets of evaluation criteria for $n$ personas in parallel.

### *3.3. Diversity-Coverage Index (DCI)*

We propose the Diversity-Coverage Index (DCI) to quantify the structured divergence of evaluation preferences across personas, independent of the human validity of the evaluation. It is formally composed of two terms derived from the distribution of optima, which, in this work, is the argmax level of dimension shifts resulting from judges:

1. Persona signal: distinctiveness of persona-specific optima, measured by normalized mutual information (NMI).
2. Distinctness: deviation from persona-agnostic baselines, measured by Earth Mover's Distance (EMD) and Jensen–Shannon divergence (JSD).

**Definition.** Let $d \in \mathscr{D}$ index a quality dimension with levels $\mathscr{L}_d = \{0, \ldots, K_d - 1\}$. Let $K_d := |\mathscr{L}_d|$, $\Delta^{K_d-1} = \{u \in \mathbb{R}_+^{K_d} : \sum_{\ell=0}^{K_d-1} u(\ell) = 1\}$, and let $L_d$ be the level r.v. on $\mathscr{L}_d$. $P$ denotes the persona random variable with weights $\pi = (\pi_a)$, where $\sum_a \pi_a = 1$ ($\Pr[P = a] = \pi_a$). For persona $a$, let $p_{a,d} \in \Delta^{K_d-1}$ be the empirical distribution of optima (argmax over cases) on dimension $d$, i.e. $p_{a,d}(\ell) = \Pr[L_d = \ell \mid P = a]$. The persona-marginal probability is defined as $q_d(\ell) = \sum_a \pi_a \, p_{a,d}(\ell) = \Pr[L_d = \ell]$. If available, $b_d \in \Delta^{K_d-1}$ serves as a persona-agnostic baseline on $\mathscr{L}_d$. Unless otherwise stated, we use equal persona weights ($\pi = 1/|\mathscr{P}|$) and set $\lambda = 0.5$. Let CDF $F_u(\ell)$ be $\sum_{j=0}^{\ell} u(j)$.
The persona signal is defined as

$$I_d := \mathrm{NMI}_d = \frac{2I(P;L_d)}{H(P)+H(L_d)} = \frac{2\,[H(q_d) - \sum_a \pi_a H(p_{a,d})]}{H(P)+H(q_d)} \in [0,1], \tag{1}$$

with joint $\Pr[P=a, L_d=\ell] = \pi_a\, p_{a,d}(\ell)$. The distinctness (persona-averaged) is

$$\bar{J}_d \coloneqq \sum_a \pi_a \,\mathrm{JSD}_{\log_2}\big(p_{a,d} \,\|\, b_d\big) \in [0,1], \quad \bar{S}_d \coloneqq \sum_a \pi_a \widetilde{W}_1\big(p_{a,d}, b_d\big) \in [0,1], \tag{2}$$

$$D_d \coloneqq \tfrac{1}{2}(\bar{J}_d + \bar{S}_d) \in [0,1], \tag{3}$$

$$\widetilde{W}_1(p,b) = W_1(p,b)/(K_d-1), \quad W_1(p,b) = \sum_{\ell=0}^{K_d-1} \big|F_p(\ell) - F_b(\ell)\big|. \tag{4}$$

The per-dimension and overall indices for $\lambda \in [0,1]$ are

$$\mathrm{DCI}_d(\lambda) = \lambda I_d + (1-\lambda) D_d \in [0,1], \tag{5}$$

$$\mathrm{DCI}(\lambda) = \tfrac{1}{|\mathscr{D}|} \sum_{d \in \mathscr{D}} \mathrm{DCI}_d(\lambda). \tag{6}$$

If no baseline exists, set $\lambda = 1$.

**Interpretation.** $I_d$ increases when personas are internally coherent yet differ across levels. $D_d$ increases when persona-specific optima deviate from generic judges, both in distributional mass (JSD) and in location/consistency along the ordered scale (EMD).

**Lemma 1 (Shuffle Sanity)** *Shuffling persona labels across i.i.d. cases removes persona effects on optimal levels, preserving the marginal $\pi$ independently of $\mathscr{L}_d$ but making $\tilde{P} \perp \mathscr{L}_d$. Thus $I(\tilde{P}; \mathscr{L}_d) = 0$ and $I_d = 0$, so $\mathrm{DCI}_d(\lambda) = (1-\lambda)D_d$. Moreover, the plug-in estimator satisfies $\widehat{\mathrm{NMI}}_d \xrightarrow{p} 0$ as $n \to \infty$, where $n$ is the total number of persona–case pairs $(P, \mathscr{L}_d)$ in the shuffled sample. Proof can be accessed on GitHub.*

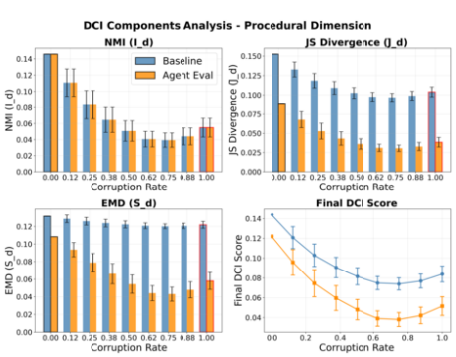
## 4. Experimental Setup

We evaluate summaries in the controlled dimension-shifted dataset using three evaluators: (1) vanilla LLM-as-a-judge (baseline), (2) AgentEval [7], and (3) PersonaMatrix. For criteria generation, we compared several reasoning LLMs (GPT-5[25], GPT-5-mini[25], Gemini-2.5-pro[26], Gemini-2.5-flash[26]) and selected GPT-5-mini. Criteria lists are available in our GitHub.

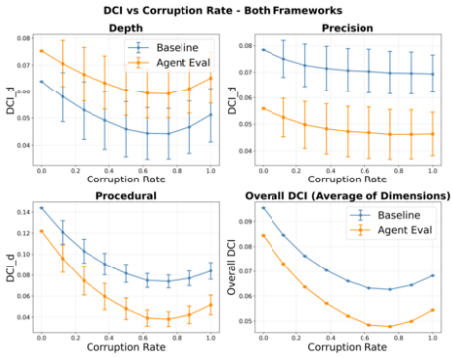
For quantification, we use public leaderboard standings which motivate this choice: GPT-5-mini (minimal-reasoning) is preferred for its stability, discriminative power (ability to separate systems in evaluation), human correlation in creative judgment, and cost-efficiency, as evidenced by its top-5 placement on EQ-Bench Judgemark-v2.1[4]. To reduce bias and improve reproducibility, we blind the Quantifier to shift levels and run three stochastic passes per summary, averaging the scores. We do not scale numerical scores, since our analysis only relies on the ordinal order of scores, converting them into argmax shift levels for each legal case.

To study the impact of persona awareness, we test the robustness of PersonaMatrix against corrupted persona labels. We conduct a ***corruption sweep*** by randomly permuting persona labels for a fraction $r \in [0,1]$ of cases. Each corruption ratio is repeated 2000 times for statistical stability. This allows us to track how DCI degrades as the persona signal is intentionally weakened. We report mean DCI scores with standard deviations as error bars (Figure 3).

[4]EQ-Bench Judgemark v2.1 Leaderboard, assessed on 2025-08-09. Archived copy: Wayback Machine snapshot (https://web.archive.org/web/20250809161133/https://eqbench.com/judgemark-v2.html).




(a) DCI Components–Procedural Dimension




(b) DCI Against Corruption Rate

**Figure 3.** (a) The changes of DCI broken down into components against the corruption ratio over Procedural dimension, (b) changes of final DCI against increasing corruption ratio. Bottom-right figure in (a) is the same as bottom-left figure in (b).

## 5. Results and Discussion

***Persona-aware judges exhibit divergent optima.*** We convert numerical scores to argmax shift levels per case, and summarize the resulting distributions in Figure 1a. In the case of ties, we use the median rounded level with the max scores as the argmax. Relative to generic judges, persona-conditioned judges exhibit shifted optima in both direction and magnitude across all three dimensions (summarized in Figure 1b). Concretely:

- Depth vs. Conciseness. Persona optima concentrate at intermediate levels (L1–L2), rather than the originals (L0), indicating that "more detail" is not universally preferred. This supports our multi-objective view of legal summarization: informative yet skim-efficient abstracts are often optimal for downstream users.
- Technical vs. Accessible. Journalists and public-facing personas tilt toward higher accessibility (L3), whereas legally oriented personas prefer more technical phrasing. Extreme accessibility (L4) is disfavored across personas, consistent with criteria that reward both clarity and retention of dispositive content.
- Procedural vs. Narrative. Most personas avoid highly narrative variants (L4), but "self-help" shows mild tolerance for L2–L4 when factual outcomes and stakes remain salient.

***Statistically significant divergence between persona-aware and generic judges.*** These shifts are not just qualitative. The $\chi^2$ tests show statistically reliable divergences from the generic judges in several persona-dimension pairs: Journalism/Media differs from Baseline on Precision $\leftrightarrow$ Accessibility ($p$=0.028) and Procedural $\leftrightarrow$ Narrative ($p$=0.001); Academic/Research differs from Baseline on Procedural $\leftrightarrow$ Narrative ($p$=0.002) and from AgentEval on Depth $\leftrightarrow$ Conciseness ($p$=0.025). These results empirically support Hypothesis H1: persona-aware judgments expose trade-offs missed by single-perspective judges.

***Operationalizing DCI: persona signal and deviation from generic judges.*** Using the argmax distributions as inputs, we compute the components of the DCI (Section 3.3). Before corruption ($r$=0), PersonaMatrix shows high persona signal (NMI) and clear deviations from persona-agnostic judges (JSD, EMD). As we gradually corrupt persona labels (shuffle rate $r \in [0, 1]$), the components decay almost monotonically across three dimensions. Figure 3a shows changes in DCI components for Procedural $\leftrightarrow$ Narrative dimension. Across three dimensions, the combined DCI drops accordingly (Figure 3b). This matches Lemma 1: shuffling destroys persona-level associations, driving NMI $\to 0$ and reducing distinctness from generic judges. These trends support Hypothesis H2: diversity signals are not artifacts of noise.

***Why does Precision $\leftrightarrow$ Accessibility decay more slowly?*** Two factors likely compress separation on this axis (Figure 1a). First, the accessibility rewrite deliberately preserves length and procedural cues, dampening gains from persona conditioning relative to the other axes. Second, readability constraints induce mid-level convergence: several rubrics reward plain phrasing and preservation of dispositive content, which tends to select L2–L3 for multiple personas, yielding flatter decay under corruption.

***High-corruption rebound is a clustering artifact.*** The small DCI uptick at high-corruption level (as $r$ increases from 0.75 to 1) shown in Figure 3b is consistent with random re-assembly of partial persona clusters during permutation. Personas with similar distribution of optima towards shift levels could be re-assigned to the same pseudo-persona, thus regaining partial persona-aware signals. For example, legal educators and litigators share similar optima patterns and may be clustered again upon near-full shuffle of persona labels.

***Linking trends to criteria: why persona-aware beats generic?*** The divergence patterns align with the learned rubrics. For instance, generic AgentEval includes "Completeness of Core Case Elements", which implicitly rewards procedural density and can bias toward L0–L1 on Depth $\leftrightarrow$ Conciseness, consistent with generic peaks at lower shift levels. In contrast, litigation-focused criteria such as "Holding clarity and citable form" and "Clarity, Conciseness, Formatting, and Navigability" value rule-like, extractable statements; this rationalizes Litigator preferences for mid-level concision and moderate accessibility (L1–L2/L3). On the narrative axis, Litigator's "Material Facts—Relevance and Completeness" and AgentEval's "Brevity vs. Necessary Detail" penalize story-heavy rewrites, explaining the broad disfavor of L4.

***Takeaways for evaluation practice:***

- Intermediate optima are the norm. Across personas, the best summaries are typically not the longest or the simplest; they sit at interior points on each axis, reinforcing that legal summarization is a multi-objective balancing problem.
- Persona lenses change the answer. Statistically significant $\chi^2$ differences and high DCI at $r$=0 show that one rubric cannot faithfully stand in for heterogeneous stakeholders.
- Corruption tests matter. The near-monotone DCI decay under label shuffling is a strong, scalable sanity check that persona signals are real rather than prompt artifacts.

## 6. Limitations and Future Work

PersonaMatrix is an LLM-driven pilot for surfacing stakeholder-conditioned preferences, but it remains an early proof of concept rather than validated stakeholder evidence. The current controlled dimension-shifted dataset is small, synthetic, and U.S.–specific, and the personas are coarse approximations of real stakeholder diversity. Consequently, the statistical evidence is illustrative and fragile, limited by sample size, lack of multiple-testing correction, and simplifying assumptions such as argmax scoring. Future studies should expand the dataset's size and domain, integrate human-annotated baselines, and test the independence of the three quality axes, since correlations among them may confound persona effects. These steps would strengthen both the robustness and generalizability of the findings.

The DCI measures systematic divergence among personas but does not capture human validity or satisfaction. Likewise, persona optima derived from agentic judges (Figure 1a) do not always align with intuitive stakeholder mappings in Table 1, emphasizing the difficulty of defining persona-specific utility. Future work should validate persona criteria and optima through stakeholder surveys that directly assess whether LLM-based evaluations align with perceived usefulness for actual users.

Rubric reliability also remains untested. Rerunning rubric generation and quantifying its stability would improve reproducibility. Some LLM-generated criteria (e.g., metadata or link checks) can be spuriously satisfied or double-count holdings, encouraging verbosity. While macro-averaging and corruption tests mitigate this, future work should detect and down-weight low-observability criteria.

Finally, the approach relies entirely on LLMs to generate summaries, define criteria, and serve as evaluators, without direct stakeholder input. For a truly stakeholder-oriented framework, human participation is essential to evaluate relevance, clarity, and utility. Establishing this external validation is the critical next step.

## 7. Conclusion

Legal summarization is inherently multi-objective, demanding evaluation frameworks that capture diverse notions of utility. PersonaMatrix addresses this by combining controlled dimension shifts, persona-conditioned criteria, and the Diversity-Coverage Index (DCI) to reveal how stakeholder perspectives diverge. Across all axes, persona-aware judges select interior optima that differ by stakeholder, whereas persona-agnostic judges collapse toward single, generic peaks. These patterns (high NMI, non-trivial JSD/EMD, and decay under label corruption) indicate systematic, non-random persona signals.

PersonaMatrix offers a practical recipe for stakeholder-sensitive evaluation in high-stakes settings. In practice: (1) one-size-fits-all judges risk mis-scoring summaries tailored to different users; (2) reporting persona-specific and generic results together exposes real trade-offs; and (3) DCI provides a compact indicator of between-persona distinctness and within-persona coherence.

Despite current scale and validation limits, PersonaMatrix shows that evaluation frameworks and, by extension, legal AI systems benefit from multiple stakeholder lenses rather than a single universal rubric. It lays groundwork for human-validated, stakeholder-aware evaluation pipelines that make AI-generated legal summaries more accurate, accessible, and socially relevant for both experts and the public.

## References


[1] Lin CY. ROUGE: A Package for Automatic Evaluation of Summaries. Text Summarization Branches Out. 2004:74-81. Available from: https://aclanthology.org/W04-1013.

[2] Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. 2002:311-8. Available from: https://aclanthology.org/P02-1040.

[3] Banerjee S, Lavie A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. 2005:65-72. Available from: https://aclanthology.org/W05-0909.

[4] Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: Evaluating Text Generation with BERT. International Conference on Learning Representations. 2020. Available from: https://openreview.net/forum?id=SkeHuCVFDr.

[5] Callison-Burch C, Osborne M, Koehn P. Re-evaluating the Role of Bleu in Machine Translation Research. In: McCarthy D, Wintner S, editors. 11th Conference of the European Chapter of the Association for Computational Linguistics. Trento, Italy: Association for Computational Linguistics; 2006. p. 249-56. Available from: https://aclanthology.org/E06-1032/.

[6] Akter M, Bansal N, Karmaker SK. Revisiting Automatic Evaluation of Extractive Summarization Task: Can We Do Better than ROUGE? In: Muresan S, Nakov P, Villavicencio A, editors. Findings of the Association for Computational Linguistics: ACL 2022. Dublin, Ireland: Association for Computational Linguistics; 2022. p. 1547-60. Available from: https://aclanthology.org/2022.findings-acl.122/.

[7] Arabzadeh N, Huo S, Mehta N, Wu Q, Wang C, Awadallah AH, et al. Assessing and Verifying Task Utility in LLM-Powered Applications. In: Al-Onaizan Y, Bansal M, Chen YN, editors. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, Florida, USA: Association for Computational Linguistics; 2024. p. 21868-88. Available from: https://aclanthology.org/2024.emnlp-main.1219/.

[8] Adler RF, Paley A, Li Zhao AL, Pack H, Servantez S, Pah AR, et al. A user-centered approach to developing an AI system analyzing US federal court data. Artificial Intelligence and Law. 2023;31(3):547-70.

[9] Kolkman D, Bex F, Narayan N, van der Put M. Justitia ex machina: The impact of an AI system on legal decision-making and discretionary authority. Big Data & Society. 2024;11(2):20539517241255101.

[10] Newman DR, Doherty U. Making the law accessible to non-lawyers: effects of different kinds of expertise on perceived usability of online legal information services. Behaviour & Information Technology. 2008;27(5):423-37. Available from: https://doi.org/10.1080/01449290601111002.

[11] Bex F. Transdisciplinary research as a way forward in AI & Law. Journal of Cross-disciplinary Research in Computational Law. 2024;2(2).

[12] Hagan M. Measuring What Matters: Developing Human-Centered Legal Q-and-A Quality Standards through Multi-Stakeholder Research. Available at SSRN 5146722. 2024.

[13] Ash E, Kesari A, Naidu S, Song L, Stammbach D. Translating Legalese: Enhancing Public Understanding of Court Opinions with Legal Summarizers. In: Proceedings of the 2024 Symposium on Computer Science and Law. CSLAW '24. New York, NY, USA: Association for Computing Machinery; 2024. p. 136–157. Available from: https://doi.org/10.1145/3614407.3643700.

[14] Liu Y, Iter D, Xu Y, Wang S, Xu R, Zhu C. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. arXiv. 2023. Available from: https://arxiv.org/abs/2303.16634.

[15] Zheng L, Chiang WL, Sheng Y, Zhuang S, Wu Z, Zhuang Y, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in neural information processing systems. 2023;36:46595-623.

[16] Blodgett SL, Cheung JCK, Liao V, Xiao Z. Human-Centered Evaluation of Language Technologies. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts. Miami, Florida, USA: Association for Computational Linguistics; 2024. p. 39-43. Available from: https://aclanthology.org/2024.emnlp-tutorials.6/.

[17] Amar R, Eagan J, Stasko J. Low-level components of analytic activity in information visualization. In: IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.; 2005. p. 111-7.

[18] Lai V, Chen C, Liao QV, Smith-Renner A, Tan C. Towards a science of human-ai decision making: a survey of empirical studies. arXiv preprint arXiv:211211471. 2021.

[19] Chang Yn, Lim Yk, Stolterman E. Personas: from theory to practices. Proceedings of the 5th Nordic Conference on Human-Computer Interaction: Building Bridges. 2008:439–442. Available from:

https://doi.org/10.1145/1463160.1463214.

[20] Salminen J, Wenyun Guan K, Jung SG, Jansen B. Use Cases for Design Personas: A Systematic Review and New Frontiers. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. CHI '22. New York, NY, USA: Association for Computing Machinery; 2022. Available from: https://doi.org/10.1145/3491102.3517589.

[21] Zhang S, Dinan E, Urbanek J, Szlam A, Kiela D, Weston J. Personalizing Dialogue Agents: I have a dog, do you have pets too? In: Gurevych I, Miyao Y, editors. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics; 2018. p. 2204-13. Available from: https://aclanthology.org/P18-1205/.

[22] Shen Z, Lo K, Yu L, Dahlberg N, Schlanger M, Downey D. Multi-LexSum: Real-World Summaries of Civil Rights Lawsuits at Multiple Granularities. CoRR. 2022;abs/2206.10883. Available from: https://doi.org/10.48550/arXiv.2206.10883.

[23] Kornilova A, Eidelman V. BillSum: A Corpus for Automatic Summarization of US Legislation. In: Wang L, Cheung JCK, Carenini G, Liu F, editors. Proceedings of the 2nd Workshop on New Frontiers in Summarization. Hong Kong, China: Association for Computational Linguistics; 2019. p. 48-56. Available from: https://aclanthology.org/D19-5406.

[24] Zhang L, Liu P, Henriksboe M, Lauvrak E, Gulla JA, Ramampiaro H. Personalsum: A user-subjective guided personalized summarization dataset for large language models. Advances in Neural Information Processing Systems. 2024;37:99333-46.

[25] OpenAI. GPT-5 System Card; 2025. Accessed 2025-09-09. Available from: https://cdn.openai.com/gpt-5-system-card.pdf.

[26] Comanici G, Bieber E, Schaekermann M, Pasupat I, Sachdeva N, Dhillon I, et al.. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities; 2025. Available from: https://arxiv.org/abs/2507.06261.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251586*

# Machine Learning to Promote Self-Composition in Conflict Resolution: The Brazilian Concil-IA Project

Isabela SABO [a], Thiago DAL PONT [b,1], Aires ROVER [a] and Jomi HUBNER [b]
[a] *Department of Law, Federal University of Santa Catarina, Florianópolis, Brazil*
[b] *Automation and Systems Department, Federal University of Santa Catarina, Florianópolis, Brazil*

**Abstract.** This paper presents the Concil-IA project, an initiative exploring how machine learning (ML) can promote self-composition in conflict resolution. The research investigates whether an ML model can predict judicial outcomes and assist parties in reaching agreements during end-procedural conciliation hearings. We collected judgments (textual documents) from the Brazilian Special Civil Court at the Federal University of Santa Catarina (JEC-UFSC), focusing on disputes in air transport services (Consumer Law) where parties require compensation value for immaterial damage. We developed an ML model following a structured pipeline, including preprocessing, data representation and labeling, model training and testing using regression algorithms, and practical application in the legal domain. The model was validated in two phases: (i) presenting predicted judgment values to parties during conciliation hearings for new, similar cases; and (ii) comparing actual judgment values with the model's predictions after the proceedings concluded. The model showed satisfactory performance with minimal deviations. Predicted values were well received by parties and their legal representatives and closely aligned with actual judgments. We conclude that an ML model can effectively support self-composition in conflict resolution, reduce procedural uncertainties, and increase the likelihood of consensual settlements in judicial disputes.

**Keywords.** Alternative Dispute Resolution, e-Justice, Machine Learning, Natural Language Processing, Consumer Law

## 1. Introduction

According to the latest edition of *Justiça em Números*, an official publication of the National Council of Justice (CNJ), the Brazilian Judiciary ended 2023 with 83.8 million ongoing lawsuits. In that year alone, around 35.3 million new lawsuits were filed. Consumer Law featured prominently in State Courts, with claims for moral and material damages ranking second and fourth, respectively, in the percentage of cases filed in 2023 [1].

[1]Corresponding Author: Thiago Dal Pont, thiagordalpont@gmail.com

Given this high level of litigation, the CNJ has invested in policies to enhance procedural management, notably alternative dispute resolution (ADR) and information technology (E-Justice). ADR was formalized as a judiciary policy through Resolution No. 125/2010 [2], aiming to promote conciliation and mediation services, train professionals, encourage cooperation for peaceful dispute resolution, and reduce judiciary congestion. Despite these efforts, the adoption of ADR has been slow, with only 12.1% of cases resolved through conciliation in 2023 [1].

E-Justice initiatives have expanded the use of artificial intelligence (AI), particularly machine learning (ML) and natural language processing (NLP), to improve decision-making. CNJ Resolution No. 332/2020 [3] established SINAPSES, the national platform for storage, supervised training, version control, distribution, and auditing of AI models, and CNJ Resolution No. 615/2025 [4] set forth guidelines for the development, use, and governance of AI-based solutions within the Judiciary. However, challenges such as data quality and labeling remain significant barriers to the effective development and implementation of such models.

In the context of AI applications for conflict resolution, Online Dispute Resolution (ODR) marks a major advancement within ADR. Initially developed for online disputes [5], ODR systems have evolved into two generations: the first, with technology playing a passive, facilitative role under human decision-making; and the second, where AI-driven tools act more autonomously, assisting or representing the parties based on prior data, while humans take on a supportive role [6]. Early ODR systems, such as Family-Winner and AssetDivider, applied game theory and heuristics to guide settlements in family disputes, while subsequent multi-agent platforms like UMCourt incorporated case-based reasoning, genetic algorithms, and behavioral analysis to suggest solutions in labor and consumer conflicts. Other initiatives, such as eJRM in Italy and TADRS in China, integrated ML and NLP to classify cases and recommend outcomes, while large-scale marketplace datasets (e.g., eBay disputes) have been used to predict resolutions with high accuracy. More recent projects include LLMediator, which integrates large language models into mediation processes, and CLAUDETTE, which applies ML and NLP to identify unfair clauses in consumer contracts [7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24].

This paper introduces the **Concil-IA project**, a research initiative that explores how ML can promote self-composition in conflict resolution.[2] The central research question is whether an ML model can predict judgment outcomes and assist parties during end-procedural conciliation hearings. The hypothesis is that such predictions can help parties make more informed decisions in conflict resolution, thereby increasing the likelihood of consensual agreements.

The structure of the paper is as follows. Section 2 presents the methodology used to develop the ML model, including data collection, preprocessing, labeling, training, testing, and application. Section 3 discusses the model validation process, where predictions were presented to parties during conciliation hearings and later compared to actual case outcomes. Section 4 provides concluding remarks and outlines future research directions.

[2]The name Concil-IA combines the Portuguese words *conciliação* (conciliation) and *Inteligência Artificial* (Artificial Intelligence), reflecting the project's goal of leveraging procedural data to facilitate conciliation and mitigate caseload in the Brazilian Judiciary.

## 2. Methodology

### 2.1. Data context and Collection

Consumer Law in Brazil, established in the early 1990s, protects individual rights and provides compensation for material and immaterial (moral) damages [25,26]. While material damage is easily quantified, immaterial damageharm to personal rights such as honor, dignity, and reputationis assessed monetarily and discretionarily, depending on factors like age, health, gender, and circumstances of the injury [27,28,29,30]. This variability makes predicting compensation suitable for machine learning, potentially aiding consensual settlements.

For this study, we collected 1,163 judgments from the Brazilian Special Civil Court at the Federal University of Santa Catarina (JEC-UFSC), issued between 2012 and 2020, focusing on disputes in air transport services. These cases involve typical issues such as flight delays, cancellations, overbooking, and baggage loss [31]. Judicial decisions on immaterial damages are context-specific due to judges' discretionary powers; therefore, our model is designed specifically for the JEC-UFSC jurisdiction, though the methodology can be adapted to other courts.

### 2.2. Machine Learning Model Development

The model was developed through a series of experiments carried out over the past five years [32,33,34,35], whose results were progressively consolidated. Its final structure consists of the four stages illustrated in Figure 1 and described below.

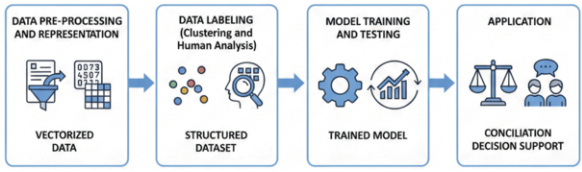



**Figure 1.** Machine Learning Workflow

#### 2.2.1. Data Preprocessing and Representation

Preprocessing was applied to prepare the legal judgments for machine learning, ensuring meaningful and manageable inputs [36]. The steps included normalization (lowercasing), tokenization (splitting text into tokens), stemming (reducing words to their stems), filtering (removal of stopwords), and the construction of bi-grams [37,38,39]. After preprocessing, the corpus was numerically represented using the *Bag of Words* (BOW) model with Term FrequencyInverse Document Frequency (TFIDF) weighting, which balances the importance of frequent versus rare terms [40].

#### 2.2.2. Data Labeling

Data labeling structures the dataset for analysis by combining expert judgment with unsupervised learning to minimize bias and reveal hidden patterns [39]. We adopted a soft clustering approach, since documents may belong to more than one category (e.g., "tem-

porary baggage loss" and "flight delay" simultaneously). Similarities between documents were measured with cosine distance [41], and clustering was performed using Hierarchical Clustering, which organizes documents into tree structures [42], and Lingo, which extracts frequent phrases to generate human-readable group labels [43].

Our experiments [34] showed that clustering-based analysis facilitated the identification of common events connecting documents, enhancing labeling reliability. The main labels were organized into two groups: (i) events generating immaterial damage, including permanent baggage loss, tampered baggage, temporary baggage loss, loss interval (days), flight cancellation, flight delay, delay interval (hours/minutes), overbooking, no-show, hindered repayment, and downgradecategorical values are represented as "yes"/"no", while intervals are numerical; and (ii) events excluding immaterial damage, covering adverse weather conditions and consumer fault, both represented as categorical "yes"/"no" values. These labels were structured in CSV format for further analysis, as shown in Table 1.

**Table 1.** Example of structuring judgment texts

| Judgment ID | Temporary baggage loss | Baggage loss interval | Flight delay | Flight delay interval | [...] | Immaterial damage value |
|---|---|---|---|---|---|---|
| 0001 | Yes | 2 | No | Not applicable | [...] | 3,000 BRL |
| 0002 | No | Not applicable | Yes | 00:10:00 | [...] | 5,000 BRL |

[...] - Other labels.

Structuring the documents with labels also enables the extraction of association rules within the judgments, revealing strong relationships between labels and the immaterial damage value. To identify these relationships, we applied the *Frequent Pattern Growth* algorithm to generate rules of the form $x \rightarrow y$, where $x$ is a set of labels and $y$ the compensation value. Each rule was evaluated using *Support* (proportion of cases satisfying both $x$ and $y$) and *Confidence* (proportion of cases with $y$ among those satisfying $x$), as shown in Equation 1.

$$a(x \rightarrow y) = \frac{Frequency(x,y)}{N} \; ; \; b(x \rightarrow y) = \frac{Frequency(x,y)}{Frequency(x)} \tag{1}$$

To enable this implementation, we categorized the numerical values *Delay interval* and *Immaterial damage* into five ranges, assisted by the quartile calculation[3], as represented in Table 2.

Table 3 presents the association rules selected with a minimum support of 1% and a minimum confidence of 70%, ordered by the highest immaterial damage range.

Overall, it is evident that the longer an airline delays a consumer's flight, the higher the monetary compensation awardedand vice versa. Notably, the temporary loss of baggage for three days typically leads to reasonable indemnities. Additionally, unilaterally cancelling the return flight due to the consumers no-show on the outbound flight is associated with mid-range compensation. Conversely, null compensation is generally linked

[3]Given a data sample, quartile of order $p$ or $pth$ quartile, denoted by $q(p)$, is a location measure where $p$ is any proportion, $0 < p < 1$, such that $100p\%$ of observations are smaller than $q(p)$. Some quartiles have particular names, for example, the $1st$ quartile or $25th$ percentile, where $q(0.25) = q1$, which means that each quartile will contain approximately 25% of the data [44].

**Table 2.** Association rules ranges

| Range | Flight delay interval | Range | Immaterial damage value |
|---|---|---|---|
| 1 | 0:00:00 - 4:00:00 | 1 | 0 BRL |
| 2 | 4:01:00 - 8:00:00 | 2 | 1 - 2,000 BRL |
| 3 | 8:01:00 - 12:00:00 | 3 | 2,001 - 5,000 BRL |
| 4 | 12:01:00 - 24:00:00 | 4 | 5,001 - 8,000 BRL |
| 5 | 24:01:00 - 72:00:00 | 5 | 8,001 - 25,000 BRL |

**Table 3.** Association rules selected

| Rule | Antecedent | Consequent | Supp. | Conf. |
|---|---|---|---|---|
| 1 | Flight delay=Yes (and)<br>Delay interval=5 | Immaterial damage=5 | 3,3% | 84,6 % |
| 2 | Flight cancellation=Yes (and)<br>No show=Yes | Immaterial damage=3 | 2,1% | 83,3% |
| 3 | Temporary baggage loss=Yes (and)<br>Loss interval (days)=3 | Immaterial damage=3 | 1,6% | 73,1% |
| 4 | Consumer fault=Yes | Immaterial damage=1 | 1,5% | 94,7% |

to the consumer's own culpable conduct, such as arriving late for the flight. In such cases, the lawsuit is likely to be dismissed as unfounded. These association rules can be used to explain potential outcomes to parties during end-procedural conciliation hearings, indicating relevant events connected to the predicted result.

### *2.2.3. Model Training and Testing*

Using the text corpus and corresponding structured data, we applied supervised learning for a regression task to predict compensation values for immaterial damage. We adopted an *Ensemble Voting* approach [45], combining diverse models including *Neural Networks*, *Bagging*, *Gradient Boosting*, and *XGBoost*, which demonstrated high predictive effectiveness [35]. The dataset was partitioned using five-fold cross-validation [46] to ensure robust performance evaluation. To enhance prediction quality, we employed: N-grams (14) during preprocessing; *Mutual Information* for feature selection [47]; outlier removal using *Isolation Forest* with 10% contamination [48]; and model complexity reduction to avoid overfitting [49,50].

We evaluated the model results based on *Root Mean Square Error (RMSE)*, *Mean Absolute Error (MAE)* and $R^2$, namely a) the average of the errors of the square differences between the predicted ($y_i$) and the actual ($\hat{y}_i$) values; b) the average of the errors when predicting the dependent variable; and c) the proportion of observed variation in the predicted values that can be explained by the regression model, where $y_i$ is the actual observed value, for instance, $i$, $\hat{y}_i$ is the predicted value, for instance, $i$, $n$ is the total number of observations, and $\bar{y}$ is the mean of the actual values, as shown in Equation 2.

$$a = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}; b = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{n}; c = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (2)$$

The model exhibited robust performance, delivering notable results. With an RMSE of 1,750 BRL, the model shows a moderate prediction error, while the MAE of 915 BRL

indicates relatively small deviations. Additionally, the $R^2$ score of 0.74 signifies that 74% of the variance in compensation values is explained by the model, underscoring its robust predictive ability and reliable fit to the data.

#### *2.2.4. Application*

Lastly, to enable a real application of the model in a legal environment in cases in which there is still no judgment, we (i) select lawsuits awaiting judgment (*new cases*), extract the labels from each of them and search in the dataset for similar situations (*past cases*); for each new case, the model receives as input a combination of structured data (the label values extracted from the new case) and unstructured data (the full-text document of a similar past case), as model input; (ii) predict the compensation value for immaterial damage, generating minimum, maximum, and average estimates for comparison with past cases, and (iii) present the predictions to the parties along with relevant association rules to provide explanations and support informed decision-making.

## 3. Validations, Results and Discussion

To comprehensively address the central research question, we conducted a thorough validation of the model across its distinct phases:

**1st** We participated in conciliation hearings at the JEC-UFSC, applying the model and presenting its predicted results to the parties involved, following the strategy of section 2.2.4. We also invited them (parties and their lawyers) to answer an anonymous questionnaire evaluating the usefulness of the information provided.

**2nd** After the judicial proceedings were concludedwhich took approximately three months to two years, depending on whether the parties appealed to higher courtswe consulted the lawsuits submitted to the conciliation hearings and verified the real results, comparing them to the ones predicted by the model.

### *3.1. Participant Observation of End-procedural Conciliation Hearings and Survey to the Parties Involved*

To facilitate our participant observation, the JEC-UFSC staff selected 13 lawsuits awaiting judgment, and the chief judge scheduled a conciliation hearing for each, assigning us as conciliators. Invitation letters were sent to all parties, and airlines (defendants) were specifically asked to bring settlement offers to initiate negotiations. Although the sample is small, the focus was on how parties received and responded to the predicted compensation values. The observation period ran from September 13 to 16, 2021, with all hearings conducted virtually. To present the predictions effectively, we prepared concise informational material covering the projects objectives, the benefits of settlement, typical compensation patterns, notable high-compensation cases, and the case-specific predicted value with associated labels and rules to guide negotiation.

After each conciliation hearing, we distributed a voluntary, semi-structured survey to plaintiffs, defendants, and lawyers to qualitatively assess the impact of the information provided. The survey combined multiple-choice (MCQs) and short-answer questions (SAQs), included select Likert-scale items, and concluded with an open-ended question to capture participants in-depth perceptions, as shown in Listing 1.

```
Q1. What is your position in the case submitted for the conciliation
    hearing? (SAQ)
        ◯ Plaintiff
        ◯ Defendant
        ◯ Plaintiff's lawyer
        ◯ Defendant's lawyer

Q2. Did you reach an agreement during the conciliation hearing? (SAQ)
        ◯ Yes
        ◯ No

Q3. If the previous answer was ''no'', what reason(s) would you
    attribute to the failure of the conciliation? (MCQ)
        ◯ Absence of one or both parties
        ◯ No plaintiff's interest
        ◯ No defendant's offer
        ◯ The defendant's offer did not meet the plaintiff's
          expectation

Q4. If the previous answer was ''yes'' or ''partial'', what reason(s)
     would you attribute to the success of the conciliation? (MCQ)
        ◯ Parties' interest
        ◯ Conciliator's data information
        ◯ Defendant's offer
        ◯ The parties waived part of the value

Q5. Was the information presented by the conciliator about local
    judgments helpful to you? (SAQ)
        ◯ Extremely helpful
        ◯ Very helpful
        ◯ Somewhat helpful
        ◯ Slightly helpful
        ◯ Not at all helpful

Q6. Were the values predicted of compensation for immaterial damage
    helpful to you? (SAQ)
        ◯ Extremely helpful
        ◯ Very helpful
        ◯ Somewhat helpful
        ◯ Slightly helpful
        ◯ Not at all helpful

Q7. Could you please measure your level of satisfaction with the
    activity provided? (SAQ)
        ◯ Very satisfied
        ◯ Somewhat satisfied
        ◯ Neither satisfied nor dissatisfied
        ◯ Somewhat dissatisfied
        ◯ Very dissatisfied

Q8. Would you like to leave a comment, critique, or positive feedback
    ? (open-ended)
```

Listing 1: Post-Conciliation Hearing Survey

The survey responses are summarized in tables, followed by our analysis. Responses to Questions 17 are presented in Table 4, while responses to Question 8 are shown in Table 5. All responses were individual, anonymous, and provided voluntarily.

Analysis of the survey responses indicates that most participants were plaintiffs and their lawyers, reflecting strong consumer interest in accessing information about immaterial damage compensation (Q1). Among the 14 conciliation hearings, only one resulted in an agreement, explaining the predominance of "no" responses regarding successful settlements (Q2).

The main obstacle to agreements was the absence of offers from the airlines before the hearings, highlighting a limitation in assessing the models practical impact (Q3). In

**Table 4.** Summary of participant responses to the post-conciliation survey (Q1-Q7)

| Question | Option | % Respondents |
|---|---|---|
| Q1. Role | Plaintiff | 5 |
| | Defendant | 1 |
| | **Plaintiff's lawyer** | **6** |
| | Defendant's lawyer | 2 |
| Q2. Agreement reached | Yes | 2 |
| | **No** | **12** |
| Q3. Reason for failure | Absence of parties | 0 |
| | No plaintiff's interest | 1 |
| | **No defendant's offer** | **10** |
| | Offer did not meet expectation | 2 |
| Q4. Reason for success | Parties' interest | 1 |
| | **Conciliator's data info** | **2** |
| | Defendant's offer | 1 |
| | Waived part of value | 0 |
| Q5. Usefulness of judgment information | Extremely helpful | 5 |
| | **Very helpful** | **7** |
| | Somewhat helpful | 1 |
| | Slightly helpful | 1 |
| | Not at all helpful | 0 |
| Q6. Usefulness of predicted compensation | Extremely helpful | 3 |
| | **Very helpful** | **8** |
| | Somewhat helpful | 1 |
| | Slightly helpful | 1 |
| | Not at all helpful | 1 |
| Q7. Satisfaction with activity | **Very satisfied** | **9** |
| | Somewhat satisfied | 3 |
| | Neither satisfied nor dissatisfied | 0 |
| | Somewhat dissatisfied | 2 |
| | Very dissatisfied | 0 |

**Table 5.** Participants in-depth perceptions (Q8)

| | Responses | Comments |
|---|---|---|
| Positive feedback | 5 | Congratulations on the initiative; This kind of information provides subsidies for the parties to decide. |
| Critique | 3 | The predicted amounts were low in relation to the circumstances of the case; The process could be held up for a long time awaiting the hearing. |
| Suggestion | 1 | Previously check with the airline if they will offer an initial value; otherwise, do not hold the hearing. |

future hearings, strategies such as sending preliminary information to the airlines legal departments could encourage more active participation. Conversely, in the hearing where an agreement was reached, participants attributed the successful outcome to the information we provided, their willingness to negotiate, and the fact that the airline presented an initial offer (Q4). Notably, our observations suggest that, in this particular case, the

**Table 6.** Comparison between predicted and real immaterial damage

| Lawsuit | Predicted Minimum | Predicted Maximum | Real immaterial damage |
|---|---|---|---|
| 1 | 4,000 BRL | 5,900 BRL | 0 BRL |
| 2 | 2,100 BRL | 4,000 BRL | 1,000 BRL |
| 3 | 6,200 BRL | 7,800 BRL | 5,000 BRL |
| 4 | 2,400 BRL | 4,300 BRL | 3,000 BRL |
| 5 | 9,000 BRL | 10,100 BRL | *Agreement in conciliation hearing* |
| 6 | 2,200 BRL | 4,300 BRL | 1,000 BRL |
| 7 | 7,500 BRL | 8,900 BRL | 6,000 BRL |
| 8 | 4,900 BRL | 6,500 BRL | 8,000 BRL |
| 9 | 4,400 BRL | 5,500 BRL | 5,000 BRL |
| 10 | 4,400 BRL | 6,000 BRL | 0 BRL |
| 11 | 5,300 BRL | 8,700 BRL | 0 BRL |
| 12 | 5,100 BRL | 5,600 BRL | 8,000 BRL |
| 13 | 4,200 BRL | 6,400 BRL | 5,000 BRL |

predicted compensation value influenced the airline (defendant) to enhance its initial offer. Consequently, the revised offer aligned more closely with the consumer's (plaintiffs) expectations, facilitating the settlement.

Overall, participants generally appreciated the information presented, including the predicted compensation amounts, although some expected higher values (Q5-Q6). This underscores the importance of providing detailed factors and rules underlying the predictions to enhance transparency and acceptance. Responses also indicate overall satisfaction with the activity and positive perceptions (Q7-Q8). While we acknowledge and value the constructive nature of the criticism and suggestions received, we respectfully disagree with the proposal to avoid holding hearings, as conciliation remains the primary goal of the JECs. This reinforces that parties initiating a lawsuit should be prepared to participate in negotiation.

Beyond its practical use, the model supports fair conflict resolution by empowering consumers to make informed decisions, whether to settle or await judgment, thereby respecting both parties' autonomy and the judicial process.

### *3.2. Comparison Between Predicted and Real Judgment Results of End-procedural Conciliation Hearing Cases*

The participation in the hearings was conducted without the presence of the presiding judge, ensuring that they did not have access to our predictions. After the judicial proceedings were concludedwhich took approximately three months to two years, depending on whether the parties appealed to higher courtswe reviewed these cases to verify their judgment results and compare them with the predictions we had provided to the parties in the conciliation hearings. Table 6 presents the predictions we reported to the parties alongside the final judgment results, considering both first- and second-degree judgments. The model was executed approximately 20 times for each case, providing a predicted minimum and maximum range for the compensation values.

Observations from the analyzed cases indicate that in lawsuits no. 4, 9, and 13, the actual compensation fell exactly within the predicted range, whereas in cases no.

2, 3, 6, 7, and 8, the real amounts were slightly above or below the predicted range. Lawsuits no. 1 and 12 showed significant deviations. Cases no. 10 and 11, related to flight cancellations and delays at the onset of the COVID-19 pandemic, were not included in the training data and involved legal uncertainty regarding force majeure. Despite this, hearings proceeded using predicted values, and the judge ultimately did not recognise immaterial damages in these cases.

To evaluate model performance, we calculated an adjusted error that only considers cases where the actual immaterial damage falls outside the predicted range: (i) zero if within the range; (ii) the difference between the predicted minimum and actual value if below the range; and (iii) the difference between the actual value and predicted maximum if above the range. The MAE was approximately 2,511.11 BRL, providing a more balanced assessment of the models predictive accuracy, while the RMSE was approximately 2,940.14 BRL. Overall, the model appears suitable for practical application, even when considering cases related to the COVID-19 pandemic (which are likely to result in the non-occurrence of immaterial damage - 0 BRL).

## 4. Conclusion

Revisiting our research question and hypothesis, we conclude that an ML model can effectively predict judicial outcomes and assist parties in conflict resolution. Furthermore, these predictions increase the likelihood of settlements and enhance the quality of conciliation hearings.

The primary contribution of this study lies in providing a structured framework for legal researchers developing ML-based solutions for judicial processes addressing Consumer Law. This framework encompasses: (i) identifying the appropriate data sources and collection methods (textual data from court decisions); (ii) extracting and structuring relevant information; (iii) selecting and applying suitable techniques to transform this information into actionable knowledge (predicting the probable case outcome); and (iv) implementing this knowledge in practice while ensuring compliance with legal procedures and preserving the parties' autonomy.

Although the model could also assist judges, its optimal application is within ADR mechanisms, where it enables parties to negotiate independently of the judiciary, reduces costs, and promotes early dispute resolution. The study faces limitations, including the opacity of ML predictions, the limited number of hearings, and the variable participation of airline representatives, which may constrain negotiation and affect model validation. Ethical aspects also merit further examination, particularly regarding user interaction in conciliation sessions.

Future work will apply Explainable AI techniques, such as SHAP, to estimate factor contributions and enhance transparency. Broader datasets across courts and domains, with greater defendant participation, will allow fairer testing of the model. Within e-Justice, we are also developing an ODR system to increase autonomy, enable humanAI collaboration with conciliators, and remove the need for researcher intervention, reinforcing AI's practical role in e-Justice and ADR.

**Acknowledgement:** This research was partly supported by FAPESC (Research and Innovation Support Foundation of Santa Catarina State) Award Number 2023TR00038 and by CAPES (*Coordenação de Aperfeiçoamento de Pessoal de Nível Superior*) - Brazil.

## References


[1] CNJ. Justiça em Números 2024. Brasília, DF; 2024. Available from: https://www.cnj.jus.br/wp-content/uploads/2024/05/justica-em-numeros-2024-v-28-05-2024.pdf.

[2] Conselho Nacional de Justiça (CNJ). Resolução nž 125 de 29 de novembro de 2010. Dispõe sobre a Política Judiciária Nacional de tratamento adequado dos conflitos de interesses no âmbito do Poder Judiciário e dá outras providências. Diário da Justiça eletrônico. 2010. Available from: https://atos.cnj.jus.br/atos/detalhar/156.

[3] Conselho Nacional de Justiça (CNJ). Resolução nž 332 de 21 de agosto de 2020. Dispõe sobre a ética, a transparência e agovernança na produção e no uso de Inteligência Artificial no Poder Judiciário e dá outras providências. Diário da Justiça eletrônico. 2020. Available from: https://atos.cnj.jus.br/atos/detalhar/3429.

[4] Conselho Nacional de Justiça (CNJ). Resolução nž 615, de 12 de março de 2025; 2025. Estabelece diretrizes para o desenvolvimento, utilização e governança de soluções de Inteligência Artificial no Poder Judiciário. https://atos.cnj.jus.br/files/original1555302025031467d4517244566.pdf.

[5] Katsh E. ODR: a look at history - a few thoughts about the present and some speculation about the future. The Hague, Netherlands: Eleven International Publishing; 2011. p. p. 21-33.

[6] Carneiro D, Novais P, Neves J. Conflict resolution and its context. Switzerland: Springer; 2014.

[7] Bellucci E, Zeleznikow J. Developing Negotiation Decision Support Systems that support mediators: a case study of the Family_Winner system. Artificial Intelligence and Law. 2005;13(2):233-71.

[8] Bellucci E. Development of a mediation tool in Family Law: AssetDivider. Conference on Legal Knowledge AND Information Systems (JURIX), 21, 2008, Florence. Amsterdam, The Netherlands: IOS Press; 2008. p. p. 88-97. Available from: https://dl.acm.org/doi/10.5555/1564008.1564019.

[9] Lodder AR, Zeleznikow J. Artificial intelligence and Online Dispute Resolution. The Netherlands: Eleven International Publishing; 2012. p. p. 73-94.

[10] Abrahams B, Bellucci E, Zeleznikow J. Incorporating fairness into development of an integrated multi-agent online dispute resolution environment. Group Decision and Negotiation. 2012;21(1):3-28.

[11] Carneiro D, Novais P, Neves J. An Agent-Based Architecture for Multifaceted Online Dispute Resolution Tools. Berlin, Heidelberg: Springer; 2011. p. p. 89-94.

[12] Carneiro D, Novais P, Andrade F, Zeleznikow J, Neves J. Using case-based reasoning and principled negotiation to provide decision support for dispute resolution. Knowledge and Information Systems. 2013;36(3):789-826.

[13] Carneiro D, Novais P, Neves J. Using genetic algorithms to create solutions for conflict resolution. Neurocomputing. 2013;109:16-26.

[14] Gomes M, Oliveira T, Carneiro D, Novais P, Neves J. Studying the effects of stress on negotiation behavior. Cybernetics and Systems. 2014;45(3):279-91.

[15] Carneiro D, Novais P, Andrade F, Zeleznikow J, Neves J. Online dispute resolution: an artificial intelligence perspective. Artificial Intelligence Review. 2014;41(2):211-40.

[16] Carneiro D, Gomes M, Costa Â, Novais P, Neves J. Enriching conflict resolution environments with the provision of context information. Expert Systems. 2017;34(5):e12049.

[17] Fersini E, Messina E, Manenti L, Bagnara G, El Jelali S, Arosio G. eMediation-Towards Smart Online Dispute Resolution. International Conference on Knowledge Management AND Information Sharing (KMIS), 2014, Rome. Setúbal, Portugal: SciTePress; 2014. p. p. 228-36. Available from: https://www.scitepress.org/Link.aspx?doi=10.5220/0005080002280236.

[18] El Jelali S, Fersini E, Messina E. Legal retrieval as support to eMediation: matching disputant's case and court decisions. Artificial Intelligence and Law. 2015;23(1):1-22.

[19] Zheng GG. China's grand design of people's smart courts. Asian Journal of Law and Society. 2020;7(3):561-82.

[20] Tsurel D, Doron M, Nus A, Dagan A, Guy I, Shahaf D. E-commerce dispute resolution prediction. Virtual Event, Ireland: ACM INTERNATIONAL CONFERENCE ON INFORMATION KNOWLEDGE MANAGEMENT (CIKM), 29, 2020, Ireland; 2020. p. 1465-74. Available from: https://dl.acm.org/doi/pdf/10.1145/3340531.3411906.

[21] Simkova N, Smutny Z. Business E-NeGotiAtion: A method using a genetic algorithm for online dispute resolution in B2B relationships. Journal of Theoretical and Applied Electronic Commerce Research. 2021;16(5):1186-216.

[22] Westermann H, Savelka J, Benyekhlef K. Llmediator: Gpt-4 assisted online dispute resolution. arXiv preprint arXiv:230716732. 2023.

[23] Lagioia F, Jabłonowska A, Liepina R, Drazewski K. AI in search of unfairness in consumer contracts: the terms of service landscape. Journal of Consumer Policy. 2022;45(3):481-536.
[24] Galassi A, Lagioia F, Jabłonowska A, Lippi M. Unfair clause detection in terms of service across multiple languages. Artificial Intelligence and Law. 2024:1-49.
[25] Lopes SM, Laniado RN. Os Juizados Especiais Cíveis de Defesa do Consumidor: a ampliação dos direitos no Estado democrático. Política & Sociedade. 2010;9(17):359-92.
[26] Brasil. Lei nž 8.078, de 11 de setembro de 1990. Dispõe sobre a proteção do consumidor e dá outras providências. Diário Oficial [da] República Federativa do Brasil. 1990. Available from: http://www.planalto.gov.br/ccivil_03/leis/l8078compilado.htm.
[27] Diniz MH. Proteção jurídica da existencialidade. Revista Eletrônica Direito e Sociedade (REDES). 2020;8(2):181.
[28] Reale M. Moral Damages in Brazilian Law. Panorama of Brazilian Law. 1992:121-7.
[29] Gonçalves CR. Responsabilidade civil. São Paulo: Saraiva; 2020.
[30] Sadiku A. Immaterial Damage and Some Types of its Compensation. Prizren Social Science Journal. 2020;4(1):50-6.
[31] Benjamim AHV. O transporte aéreo e o código de defesa do consumidor. Revista de Direito do Consumidor. 2015;100(24):23-37.
[32] Sabo IC, Dal Pont TR, Rover AJ, Hubner J. Classificação de sentenças de Juizado Especial Cível utilizando aprendizado de máquina. Revista Democracia Digital e Governo Eletrônico. 2019;1(18):94-106.
[33] Dal Pont TR, Sabo IC, Hübner JF, Rover AJ. Impact of Text Specificity and Size on Word Embeddings Performance: An Empirical Evaluation in Brazilian Legal Domain. In: Cerri R, Prati RC, editors. Intelligent Systems. Cham: Springer International Publishing; 2020. p. 521-35.
[34] Sabo IC, Dal Pont TR, Wilton PEV, Rover AJ, Hübner JF. Clustering of Brazilian legal judgments about failures in air transport service: an evaluation of different approaches. Artificial Intelligence and Law. 2021:1-37.
[35] Dal Pont TR, Sabo IC, Hübner JF, Rover AJ. Regression applied to legal judgments to predict compensation for immaterial damage. PeerJ Computer Science. 2023 Mar;9:e1225.
[36] García S, Luengo J, Herrera F. Data preprocessing in data mining. vol. 72. Switzerland: Springer International Publishing; 2015.
[37] Jurafsky D, Martin JH. Speech and Language Processing: an introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 1st ed. USA: Prentice Hall PTR; 2000.
[38] Jivani AG, et al. A comparative study of stemming algorithms. Int J Comp Tech Appl. 2011;2(6):1930-8.
[39] Kotu V, Deshpande B. Data Science. 2nd ed. Cambridge, MA: Morgan Kaufmann (Elsevier Science); 2019.
[40] Hu X, Liu H. Text analytics in social media. New York: Springer; 2012. p. 385-414.
[41] Cichosz P. Data mining algorithms: explained using R. Chichester: John Wiley & Sons; 2015.
[42] Aggarwal CC. Machine learning for text. Cham: Springer; 2018.
[43] Osiński S, Stefanowski J, Weiss D. Lingo: Search results clustering algorithm based on singular value decomposition. Berlin, Heidelberg: Springer; 2004. p. p. 359-68.
[44] Bussab WdO, Morettin PA. Estatística básica. São Paulo: Saraiva; 2010. .
[45] Mendes-Moreira J, Soares C, Jorge AM, Sousa JFd. Ensemble approaches for regression: a survey. Acm computing surveys (csur). 2012;45(1):1-40.
[46] Wong TT. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recognition. 2015;48(9):2839-46.
[47] Cover TM, Thomas JA. Elements of information theory second edition solutions to problems. Internet Access. 2006:19-20.
[48] Liu FT, Ting KM, Zhou ZH. Isolation forest. Pisa, Italy: IEEE INTERNATIONAL CONFERENCE ON DATA MINING (ICDM), 8, 2008, Pisa; 2008. p. p. 413-22. Available from: https://ieeexplore.ieee.org/document/4781136.
[49] Karystinos GN, Pados DA. On overfitting, generalization, and randomly expanded training sets. IEEE Transactions on Neural Networks. 2000;11(5):1050-7.
[50] Liu R, Gillies DF. Overfitting in linear feature extraction for classification of high-dimensional image data. Pattern Recognition. 2016;53:73-86.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251587*

# Validity Diagrams

Diogo SASDELLI [a,1] and Bianca STEFFES [b]
[a] *Department for E-Governance, University for Continuing Education Krems, Austria*
[b] *Saarland University, Saarland Informatics Campus, Germany*
ORCiD ID: Diogo Sasdelli https://orcid.org/0000-0002-6504-9812, Bianca Steffes https://orcid.org/0009-0001-9784-8942

**Abstract.** The legal visualisation method of normative diagrams is a powerful tool to visualise legal norms. However, a major limitation of normative diagrams in their original form is their sole focus on visualising norms regulating human behaviour. The paper at hand proposes to extend the framework of normative diagrams by adding *validity diagrams*, which may illustrate the validity of the behaviour norms already incorporated in normative diagrams in their original form. This extension further enhances the capabilities of normative diagrams and allows for a thorough visualisation, e.g. of conditional norms and legal *if... then...* structures. By providing different visualisations for different kinds of normative conditions, the improved method of normative diagrams presented here also contributes to clarifying the differences between various types of conditional norms.

**Keywords.** Normative Diagrams, Legal Visualisation, Commitment, Conditional Norm, Hypothetical Norm, If-Then-Structure, Validity Diagrams, ViNo

## 1. Introduction

Legal visualisation methods employ human-oriented, (semi-) formal and mainly pictorial (instead of verbal) representations to facilitate the understanding of complex legal concepts and phenomena [1], improving, e.g., the understanding of law by laymen, law students and even legal experts [2,3,4,5,6]. More formally structured methods can also function as an intermediary between natural language and proper formalisations (e.g., through logical formalisms or programming languages), thus also improving interdisciplinary communication [7,8,9]. An example of a formally structured legal visualisation method is the method of normative diagrams (or *Trivisonno-diagrams*), which was first developed for didactic purposes by legal philosopher *Alexandre T. G. Trivisonno* [5]. The method is particularly flexible, being able to provide elegant visualisations of several legal phenomena such as *gaps*, *antinomies* and different forms of legal arguments. Notwithstanding, the method, as originally presented [5], suffers from a crucial limitation: it can only visualise behaviour norms (prohibitions, obligations, permissions) and is thus unable to adequately represent aspects related to the *validity* of norms (e.g., the nullity of a contract or the validity conditions of an administrative act).

The paper at hand discusses this limitation and proposes a solution based on extending the diagrams' underlying legal theory to also include so-called *validity diagrams*. To

[1] Corresponding Author: Diogo Sasdelli, diogo.sasdelli@donau-uni.ac.at.

this end, Sect. 2 first provides a brief introduction to the method of normative diagrams. Sect. 3 then discusses the above-mentioned validity limitation, providing an example containing three legal issues, of which two cannot be adequately visualised using the method in its original form. Sect. 4 presents the proposed validity diagrams and demonstrates how they can solve the validity limitation, while also being based on the same legal-positivist theoretical frameworks underlying normative diagrams in their original form. Sect. 5 then discusses how the interplay between validity and normative diagrams contributes to clarifying the role of legal validity with respect to the notion of conditional (or hypothetical) norms. Finally, Sect. 6 concludes and refers to future work.

## 2. Normative Diagrams

Normative diagrams are based on an analytical concept of legal norms particularly popular among modern legal positivists such as *Norberto Bobbio* [10, p. 214] and *Hans Kelsen* [11, p. 116-119]. According to this view, legal norms either *prohibit*, *obligate* or *permit* (1) specific *types of actions* performed by (2) specific *actors* in (3) specific *locations* at (or during) (4) a specific point (or period) in *time*. Hence, norms can be defined as being normative (deontic) valuations (*prohibition*, *permission*, *obligation*) ranging over a four-dimensional (*object*, *space*, *subject*, *time*) validity domain.

Consider, e.g., a norm $\phi$ in a rental agreement prohibiting tenants from playing the piano in the house between 10 pm and 7 am. By respectively defining the action-type of *playing the piano* as being delimited by the interval $[o_1, o_2]$ (objective dimension); the locations in the house by the interval $[r_1, r_2]$ (spatial dimension), the tenants by $[s_1, s_2]$ (subjective dimension), and the time between 10 pm and 7 am by $[t_1, t_2]$ (temporal dimension), the validity domain $U_\phi$ of $\phi$ can defined by the expression:

$$U_\phi = \{ <o_w, r_x, s_y, t_z> \mid o_w \in [o_1, o_2] \wedge r_x \in [r_1, r_2] \wedge s_y \in [s_1, s_2] \wedge t_z \in [t_1, t_2] \} \quad (1)$$

In this sense, every ordered quadruple $< o, r, s, t >$ can be considered as a *case*; and $U_\phi$ is the set of all cases that can be subsumed under $\phi$, i.e., in the example at hand, cases in which a tenant plays the piano in the house between 10 pm and 7 am.

By mapping each of the four dimensions to a respective coordinate of a coordinate system and assuming an adequate (possibly arbitrary) ordering, norms can be visualised as *n-orthotopes* ($1 \leq n \leq 4$), i.e., as lines, rectangles, orthogonal parallelepipeds or hyperrectangles, depending on how many coordinates are simultaneously displayed. Cases, in their turn, can be represented as points. Since four-dimensional concepts are unintuitive, it is usually better to visualise a norm's validity domain in two or more figures, as shown in Fig. 1 (cf. also [12]). Furthermore, if a norm's validity domain extends over the entirety of a dimension, this dimension can be omitted. Since legal norms are often general, being applicable at all times to all legal subjects in the entirety of a given jurisdiction, it is rarely necessary to represent more than one or two dimensions at a time.

Normative diagrams are a flexible visualisation method that can be employed to visualise different types of legal arguments such as legal subsumption, analogy, and *e contrario* argumentation as well as legal phenomena such as gaps, conflicts between legal norms [5] and more nuanced aspects of normativity like supererogation [8]. Analogical reasoning, for example, is visualised as an extension of a given norm's validity domain in order to reach a case close to it, using geometric proximity to represent similarity. This

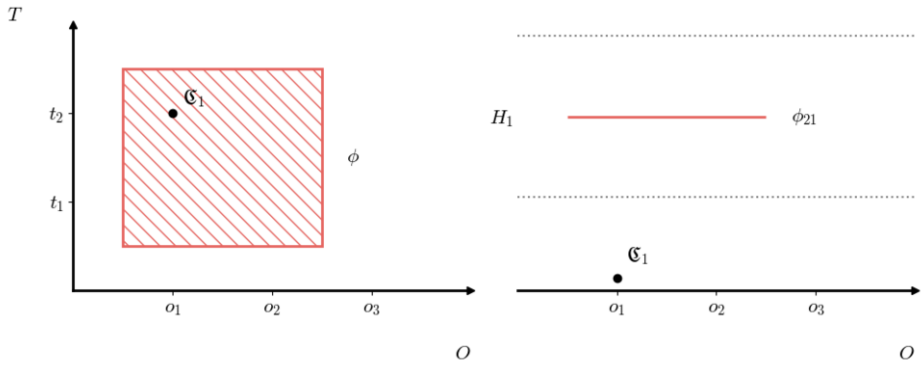



**Figure 1.** Normative diagrams for the norm $\phi$. $\mathfrak{C}_1$ represents a case in which a tenant plays piano in the house between 10 pm and 7 am, violating $\phi$.

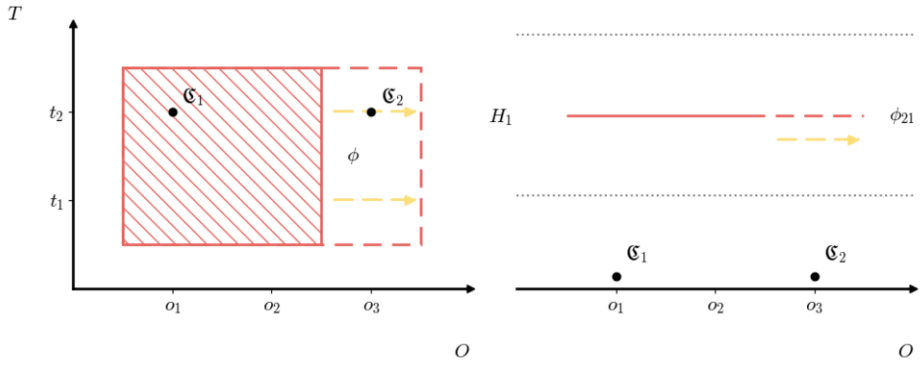



**Figure 2.** An example of the visualisation of analogical reasoning through normative diagrams.

is visualised in Fig. 2, showing how norm $\phi$ could be applied, by analogy, to a case $\mathfrak{C}_2$ in which a tenant plays the flute in the house between 10 pm and 7 am. The diagrams are particularly compatible with first-order formalisms such as PROLEG [8]. An open access tool for generating normative diagrams called ViNo [13,14] is already available and being continuously improved.[2] The current version (ViVa ViNo) includes an implementation of the validity diagrams presented here.

## 3. Validity as a Limitation of Normative Diagrams

Despite their flexibility, normative diagrams suffer from a fundamental limitation: they are only capable of representing norms[3] in a rather strict sense, i.e., norms that are directly aimed at human behaviour, establishing that certain actions are prohibited, permit-

[2] The code can be found at https://github.com/bs-000/norm-visuals. The repository also contains a Windows executable file to provide an easy access for technical laymen. All figures in the paper at hand have been created using ViNo.

[3] Here, *norms* are understood as being the *result* of interpreting legal sources (statutes, contracts etc.) [15].

ted or obligatory. These norms roughly correspond to what H. L. A. Hart calls *primary rules* [16, p. 81] or to what Hans Kelsen calls independent norms (*selbständige Normen* [17, p. 52]), especially with respect to what he calls the normative functions of obligation, prohibition and permission [11, p. 76-81]. However, legal systems also contain other kinds of norms (in a somewhat wider sense), namely those that establish the conditions under which new norms can be created, altered or eliminated (derogated), i.e., those norms which Hart calls *secondary rules* [16, p. 81] and Kelsen calls dependent norms (*unselbständige Normen* [17, p. 52]), especially with respect to the normative functions of empowering (*Ermächtigen*) and derogating (*Derogieren*) [11, p. 82-88]. Thus, the theoretical framework underlying normative diagrams ignores all those formal or procedural aspects that determine the validity of statutes, administrative acts and contracts. For example, a non-valid contract would imply non-valid obligations, which, as such, would simply lead to an empty representation (i.e., no representation at all) within normative diagrams [8].

In many cases, especially within private law, the legal analysis is focussed precisely on whether procedural formalities or other formal requirements have been observed, resulting in either the validity or in the nullity of the respective legal acts, transactions etc. Since normative diagrams are only able to represent cases involving behaviour norms (*prohibitions*, *obligations*, *permissions*), they are unable to provide visualisations for such cases focussing on legal validity, which therefore constitutes a major limitation of normative diagrams as a legal visualisation method. This limitation can be illustrated through the following example.

**Example.** Titius *(the tenant), a young aspiring pianist studying at the local music faculty in a small town in Austria, signed a rental agreement with* Mevius *(the landlord). The rental unit is a small apartment in the upper floor of* Mevius*' house, which lies relatively isolated in the town's outskirts. The ground floor contains another apartment, in which* Mevius *himself lives, and a small shop which is currently occupied by a local pharmacy. In the rental agreement, they included a clause ($\phi$) according to which* Titius *would not be allowed to play the piano between 10 pm and 7 am. A year later, with difficult exams approaching,* Titius *sent* Mevius *an e-mail asking for permission to practice the piano also at night.* Mevius *replied via e-mail that, to prepare for his exams,* Titius *would be exceptionally allowed to play the piano until 2 am once a week in the night from Monday to Tuesday (let this exception be called $\psi$). After this,* Titius *started practising until 2 am on Mondays at the end of every semester. 18 months (i.e., three semesters) later, as* Titius *did this once again,* Mevius *claimed* Titius *had broken the agreement and terminated the contract with effect at the end of the following month, which eventually led to an eviction procedure. In court,* Mevius *argued that he granted the exception only once, and* Titius *ought to have asked for permission once more before starting to practise at night again. Moreover, he argued that the exception was not contractually binding, because it lacked a proper written form (being only an e-mail). Finally,* Mevius *argued that, since the house was so small, containing only two apartments, the rental agreement would be subject not to the Austrian rental law act (*Mietrechtsgesetz *– MRG), but to the general civil code (*Allgemeines Bürgerliches Gesetzbuch *– ABGB), which would allow him to terminate the contract without a particular reason.* Titius *contested these claims and argued, in particular, that the fact that* Mevius *only complained after 18 months (i.e., after* Titius *had been using the exception for three semesters) indicates that the exception was generally valid for whenever* Titius *would be preparing for exams.*

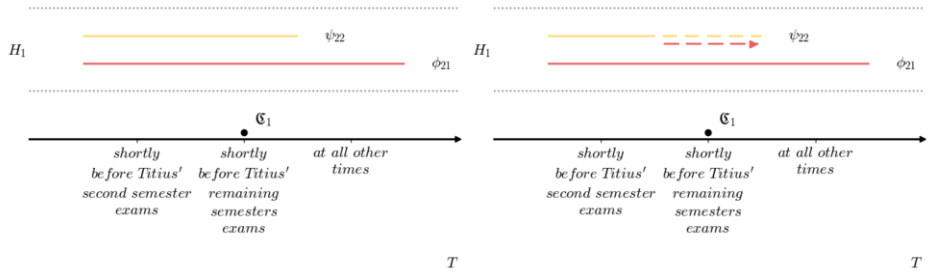



**Figure 3.** Diagrams illustrating *Mevius*' and *Titius*' disagreement on the extent of $\psi$, i.e., whether it does encompass *Titius*' latest practising $\mathfrak{C}_1$ (on the left) or not (on the right).

Evaluating *Mevius*' arguments requires analysing three distinct legal issues:

1. The *extent* of the exception granted by *Mevius*, i.e., whether it was granted only once or whenever *Titius* would be preparing for approaching exams.
2. The legal validity of this exception, i.e., whether it was contractually binding or not.
3. The legitimacy of *Mevius'* act of terminating the contract, i.e., whether it had the legal effect of derogating the contract or not.

The first issue can be represented as a disagreement concerning the temporal extension of $\psi$ (i.e., the exception to $\phi$). This can be adequately illustrated through normative diagrams as shown in Fig. 3. The other two issues, however, go beyond their limits. The second issue basically asks whether $\psi$ legally *exists* or not. Normative diagrams could, at most, try to visualise this by including or not including $\psi$ in a representation. This, however, would not contribute to visualising the actual legal matter being discussed, i.e., whether the fact that the exception was granted solely via e-mail affects its legal validity or not. Finally, the third and last issue does not directly affect $\psi$. In fact, it is not even related to any behaviour norm (prohibition, permission obligation). Instead, *Mevius*' attempt to terminate the contract aims at suppressing *Titius*' rights as a tenant. This is a necessary pre-requisite for the legitimacy of the eviction procedure, because the rightful tenant, of course, may not be evicted. Hence, this third issue lies completely outside the scope of normative diagrams.

## 4. Solution through Validity Diagrams

This limitation of normative diagrams as a legal visualisation method can be overcome by complementing their underlying theoretical framework in order to also account for non-behaviour norms, i.e. Hart's *secondary rules* or Kelsen's non-autonomous norms. These norms are not concerned with human behaviour, but with the *validity* of other norms, i.e., with their status (or, as Kelsen puts it, their *specific existence* [11, p. 130]) as legal norms. For this reason, it seems adequate to call diagrams representing these norms *validity diagrams*.

As already mentioned above, validity diagrams would be concerned not with prohibitions, obligations and permissions, but with *empowerments* – i.e., the creation of legal norms – and with *derogations* – i.e., the suppression or elimination of legal norms.

Following Kelsen, a *norm* can be defined as being the "meaning of an act of will" (*Sinn eines Willensaktes*), i.e., *grosso modo*, an act through which human will is manifested or expressed [11, p. 2]. Kelsen gives the example of a group of people entering a room, then debating and voting on a specific subject [18, p. 16-17]. These are acts of will of the individuals involved; their legal meaning is, for example, the approval of a bill, i.e., a legal norm. Thus, norms are created – or *mutatis mutandis* suppressed – through human *acts* (of will). Within Kelsen's theoretical framework, an act of will shall have a legal meaning (i.e., shall correspond to a *legal* norm) if this meaning is given to it by another legal norm, i.e., roughly speaking, if this act conforms to a legal norm, establishing, e.g., the procedures that have to be followed by parliament when approving bills [19].

Based on this theoretical framework, it is possible to design *validity diagrams* in a fairly analogous way to how *normative diagrams* are structured. In particular, since norms are created or suppressed by acts, validity diagrams can be built on the same four dimensions underlying normative diagrams: an act of will can also be analysed in *object*, *space*, *subject* and *time*. Instead of prohibitions, obligations and permissions, validity diagrams contain two categories of legal norms: *empowerments* and *derogations*; jointly, these two categories can be called *authorisations*. Furthermore, to account for the legislative practice of establishing general rules followed by specific exceptions, it seems reasonable to distinguish between positive and negative forms of authorisation (i.e., of empowerments or of derogations). For example, *private autonomy* can be seen as a general positive empowerment to all people with legal capacity to establish contractual relationships with one another. This authorisation is limited by negative empowerments which exclude, e.g., the possibility of dealing with illegal items like weapons or drugs. Like prohibitions, obligations and permissions, authorisations also stretch over a four-dimensional validity domain – cf. equation (1), above. An ordered quadruple $< o, r, s, t >$, i.e., a case, corresponds to an act of will. If this act falls under an authorisation (e.g., by subsumption or by analogy), the act is *authorised*, and it produces the respective legal effect, i.e., creating (empowerment) or suppressing (derogation) a respective legal norm. Finally, as Kelsen notes, authorisations are *non-autonomous* norms (in Hart's terminology: *secondary rules*), meaning that they are always "attached" to other norms, to whose validity they refer to and which can be either other authorisations or *autonomous* norms (*primary rules* in Hart's terminology), i.e., prohibitions, obligations or permissions. One can therefore describe *validity chains* among legal norms. For example, a command given to somebody by their superior has its validity conditioned by their respective act of will (their act of commanding) falling under an authorisation (e.g., the employment contract), which in turn has its validity conditioned by the respective act of will (our signing of the contract) falling under another authorisation (e.g., labour law) etc. Following Kelsen, the ultimate source of all legal validity is a hypothetical so-called *basic norm* [11, p. 205]. Thus, each norm is part of a validity chain starting with the *basic norm* and ending with an autonomous norm (prohibition, obligation, permission) [18, p. 75-76]. To account for this, it is important that authorisations shown on validity diagrams include a reference (e.g., through an index) to the norms to whose validity they refer to.

Complementing the original framework of normative diagrams with validity diagrams as described above enables the visualisation of legal scenarios focussing not only on behaviour norms, but also on the general validity of legal norms, thus closing the gap identified in Sect. 3. This can be illustrated by providing visualisations to the second

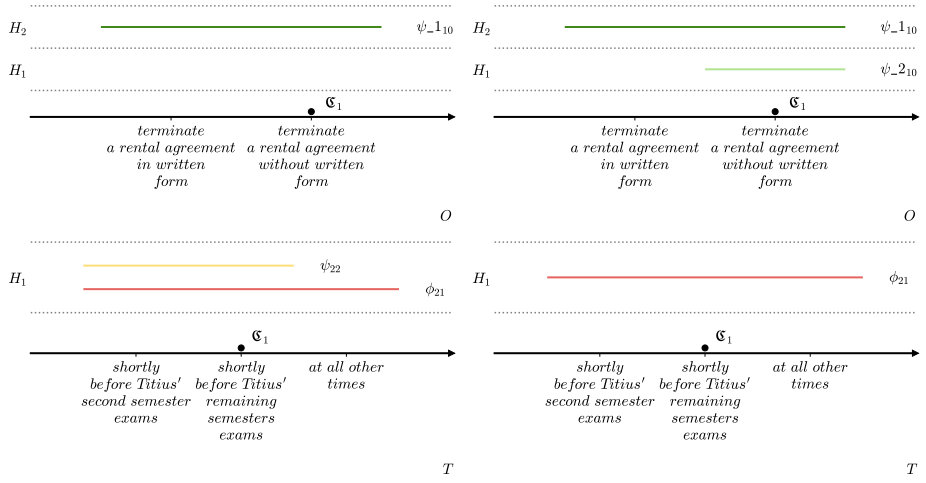



**Figure 4.** On the right: visualisation of *Mevius*' argument against the validity of $\psi$ – if there would be an exception $\psi_2$ to private autonomy ($\psi_1$), his act of will $\mathfrak{C}_1$ would not have been authorised, so that $\psi$ would not be legally valid (i.e., would not legally exist). On the left: situation without the exception $\psi_2$ – the act $\mathfrak{C}_1$ is authorised by $\psi_1$, leading to the validity (existence) of $\psi$.

and third legal issues involved in the Example (above), for which normative diagrams, in their original form, could not provide an adequate visualisation.

As already mentioned above, the first issue of the Example can be seen as a disagreement concerning the extension of the permission norm $\psi$ on the temporal dimension. This can be visualised with classic normative diagrams as shown above in Fig. 3. Arguably, *Mevius*' argumentation with respect to this issue is not convincing: both the literal wording of the exception he granted to *Titius* and his behaviour indicate that the temporal extension of $\psi$ is not limited only to the first semester in which it was granted. Thus, *Mevius* would be probably better off by contesting the validity of $\psi$ altogether, which leads to the second legal issue in the Example.

The second issue, i.e., whether the exception granted by *Mevius* was contractually binding or not, can be reduced to the question of whether the corresponding act of will, i.e., *Mevius* answering *Titius*' request via e-mail, was authorised or not; in other words, whether his act falls under a corresponding empowering norm. Overall, the principle of *private autonomy* ($\psi_1$) – i.e., the general positive empowerment to establish contractual relationships – would grant his act the required authorisation, leading to the legal validity of $\psi$. However, against this view, *Mevius* tries to argue that, in the case at hand, the lack of a proper written form would constitute an exception to general private autonomy; in other words, *Mevius* claims that there is a negative empowerment ($\psi_2$) requiring a specific written form for the act of will related to $\psi$, which would not be fulfilled by a mere e-mail. These claims are visualised in Fig. 4. Again, *Mevius*' arguments are not convincing: following § 883 of the Austrian general civil code (ABGB), contracts generally do not require a written form – except when legally required. However, neither the ABGB itself nor the Austrian rental law act (MRG) generally require a written form for rental agreements, let alone for specific agreements concerning the usage of the object. Indeed, the MRG only requires a written form for agreements concerning the establishment of

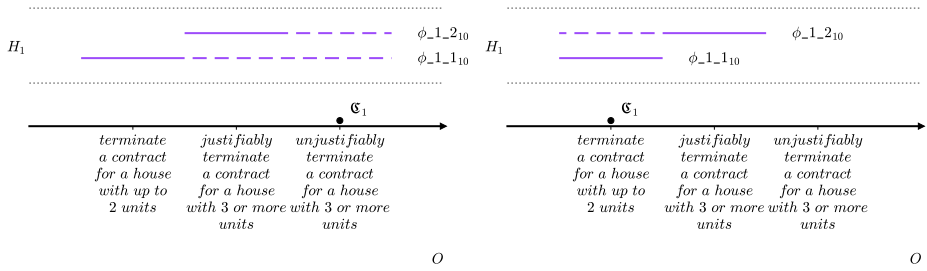



**Figure 5.** Visualisation of the third legal issue in the Example. On the right: *Mevius*' act of terminating the contract ($\mathfrak{C}_1$) falls under the ABGB and is authorised through $\phi_1_1$; on the left: his act falls under the MRG, but is not authorised through $\phi_1_2$ due to it lacking an adequate reason.

a specific duration for the contract. Hence, it follows that $\psi$ is legally valid, meaning that *Titius* did not break the agreement. Thus, *Mevius*' only remaining option is to try to argue that he can terminate the contract even without *Titius* breaking the agreement, which leads to the third legal issue in the Example.

The third issue, i.e., the legitimacy of *Mevius*' act of terminating the contract, amounts to whether this act falls under a corresponding derogative authorisation or not. *Mevius* tries to argue that the fact that the house only contains two apartments makes the rental agreement subject to the ABGB instead of to the MRG, the main difference being that only the ABGB allows the landlord to terminate the contract without a particular reason. The question here is therefore whether *Mevius*' act of terminating the contract ($\mathfrak{C}_1$) falls under the ABGB or under the MRG – or rather, under their respective derogative authorisations $\phi_1_1$ and $\phi_1_2$. This is visualised in Fig. 5. Following § 1(2)(5) MRG, buildings containing at most two homes or stores are excluded from its scope. Since *Mevius*' house contains not only two apartments, but also a small pharmacy, his claim does not hold. Thus, his act ($\mathfrak{C}_1$) falls under the MRG, which, however, requires a proper reason for terminating rental agreements. Thus, *Mevius*' act is not authorised and the contract is therefore not terminated, meaning that *Titius* is still the lawful tenant and may not be evicted.

## 5. Discussion: Validity and Conditional Norms

The main purpose of normative diagrams is to provide relatively intuitive visualisations of complex legal concepts and phenomena. Their primary application is within legal education, especially for familiarising law students with formal methods of knowledge representation. Furthermore, they can also facilitate interdisciplinary communication among lawyers, computer scientists and engineers, e.g., within projects aiming at formalising legal norms [9]. Sometimes, however, by offering a new perspective to existing issues, they might also promote conceptual clarity and, thus, lead to more general theoretical advancements [5]. This also seems to be the case concerning the concept of the *conditional norm* and its relation to the notion of legal validity: as discussed below, normative diagrams can highlight important distinctions between different kinds of normative conditions which might be overseen if only textual representations are considered.

It is widely agreed that legal norms are conditional or *hypothetical* in nature (cf. [11, p. 15-17], [20]). This is often represented by means of an *if... then...* structure, e.g.: *if certain conditions are met, then respective legal consequences ought to follow*. The usage of conditional structures as a natural way to represent legal norms has a very long tradition. For example, many of the precepts of the ancient Roman *Laws of the Twelve Tables* (5th Century BC) were already formulated in this way (e.g., in Table I, the provision "*si ambo praesentes, solis occasus suprema tempestas esto*", i.e. *if both parties are present, the sunset ought to be the time limit for the proceedings*). Within legal theory, conceptualising legal norms as having a hypothetical structure became particularly popular since the 20th Century – to a large extent through Kelsen's influence, who explicitly stressed the hypothetical nature of legal norms ("*wenn A ist, so soll B sein*") [18, p. 23].

The conception of norms as *if... then...* sentences was well received by approaches employing formal methods to represent and build models of law and legal reasoning, especially within deontic logic, leading to a long philosophical discussion concerning the adequate formalisation of the conditional norm (which von Wright called "commitment" [21]) and to the development of various systems of so-called *dyadic deontic logic* [22, 23]. This conditional structure is particularly convenient when attempting to formalise legal norms as computer programmes, which likewise often take the form of *if... then...* rules or so-called *horn clauses* (cf., e.g., [24,25]). However, as already made clear, e.g., by the above-mentioned long philosophical debate on the adequate formalisation of the conditional norm, the intuitive sense of the *if... then...* sentences traditionally employed by jurists in natural language to formulate and explain legal norms is relatively vague and ambiguous (cf. below), and thus often diverges from the strict meaning of *if... then...* sentences in formal languages, e.g., in systems of logic or in programming languages. To account for this semantic divergence, it would be necessary to conceptualise the various types of natural language *if... then...* sentences within law and legal reasoning and to then design corresponding formal language correlates to model them, instead of using one and the same formal *if... then...* structure to model all legal norms, as is often proposed in systems of deontic logic or similar models of legal reasoning (cf., e.g., [26,27]).

A closer analysis reveals that the legal theoretical notion of the conditional or hypothetical norm may refer to various different aspects of (legal) normativity. Lopes, for example, notices that conceptualising legal norms as hypothetical imperatives may imply understanding norms either (1) as means-ends-relations (i.e., in the sense of Kant [28, GMS IV 414-415] or (2) as a twofold proposition involving both (a) the relationship between the commands or imperatives as set by an authority (the *imperator*) and the respective duties of the legal subjects; and (b) the view that such commands and duties are applied conditionally, i.e., only to hypothetical behaviours or circumstances [29]. Following Kelsen, one can set aside sense (1) – means-ends-relations – as being descriptive, rather than normative [11, p. 12] (cf. also [30, p. 5]). The remaining senses (a) and (b) of conditional norms and the difference between them can be adequately visualised by our extended normative diagrams.

In sense (a), the conditionality of a norm refers to its *validity condition*, i.e., to the way how the norm was set by the respective authority (*imperator*). This is precisely the sense that motivated the introduction of validity diagrams, as discussed above. Thus, an example of conditionality in sense (a) is how norm $\psi$ (the exception allowing *Titius* to practice the piano at night) has its validity conditioned by the actions of *Mevius* (the imperator) under $\psi_1$ (private autonomy), as shown above in Fig. 4.

In sense (b), the conditionality of a norm refers to the specific set of conditions that trigger its legal effect, i.e., its *violation* (or fulfilment) *condition* or, in other words, its validity domain. Thus, sense (b) of conditionality would be represented by normative diagrams in their classic form. An example of conditionality in this sense is the extension of $\psi$'s validity domain, i.e., whether it is valid only for the first semester it was granted or whenever *Titius* would be preparing for exams, as shown above in Fig 3.

Normative diagrams (with the additions proposed in the paper at hand) provide different visualisations for these different senses of conditionality (i.e., senses (a) and (b)), thus adequately distinguishing between them. Notwithstanding, the framework described here still has an important limitation: since its structure is optimised for representing specific actions performed by specific actors in a certain place at a given time, it cannot adequately represent situations in which part of the normative condition involves *another* action being performed by potentially different actors etc. This can occur both with respect to validity conditions (sense (a)) or to violation/fulfilment conditions (sense (b)) of a norm. For sense (a), one can easily find examples in legislative procedure, which often involves various acts from different authorities (e.g. votings in various parliament chambers, signing by the president etc.) before a law is properly enacted. For sense (b), one can find examples in more complex normative situations such as *contrary-to-duty* [31] scenarios or *compensatory* norms [27], in which norms are set for situations in which another norm has already been violated. This limitation of the current framework of normative diagrams points to the particular nature of these complex normative conditions (that are often associated with so-called *paradoxes of deontic logic* [32]), which thus should be adequately accounted for by formal models of legal reasoning.

## 6. Conclusion and Future Work

The original framework for normative diagrams [5] is flexible, but suffers from the crucial limitation of not being able to adequately represent aspects related to the validity of legal norms. The paper at hand addressed this gap by expanding this original framework – while also coherently considering its original legal-theoretical foundations – to also include *validity diagrams*. By discussing the different ways of visualising various kinds of normative conditions through normative diagrams, the paper at hand also contributes to clarifying the various senses of conditionality involved in the notion of the conditional (or hypothetical) norm, stressing the importance of addressing the differences between these senses when developing formal models of legal reasoning.

The main topic for future work involves addressing the still remaining limitation of normative diagrams concerning the visualisation of normative conditions involving other actions being performed. One approach to adequately visualise these more complex conditions would be to expand the current framework of normative diagrams by allowing for a norm to be represented by several diagrams in parallel – one for each action relevant to the norm represented –, and define the respective validity and violation/fulfilment conditions on the basis of all diagrams. Thus, the current framework would correspond to a special case in which only one diagram is needed to represent a norm. Further topics for future work include the development of comprehensive semantics for normative diagrams, including investigating the relationship between this semantics and the model theory of established systems of deontic logic.

## References


[1] Cyras V, Lachmayer F. Essays on the Visualisation of Legal Informatics. Springer; 2023. Available from: https://doi.org/10.1007/978-3-031-27957-7.

[2] Röhl KF, Ulbrich S. Recht anschaulich: Visualisierung in der Juristenausbildung. Edition Medienpraxis. Halem; 2007.

[3] Kessel C. Kinder finden das Gesetz: Transparentes Recht für Kinder – Child-friendly Justice dank Visualisierung. Jusletter IT. 2014.

[4] Carnein M, Quiring E, Haack A, Möhring A, Becker J. Laiengerechte Erzeugung von 3D-Animationen am Beispiel von textuellen Unfallbeschreibungen. Jusletter IT. 2014.

[5] Sasdelli D, Trivisonno A. Normative Diagrams as a Tool for Representing Legal Systems. The Review of Socionetwork Strategies. 2023 08;17:217-31. Available from: https://doi.org/10.1007/s12626-023-00144-0.

[6] Muff F, Fill HG. Towards Embedding Legal Visualizations in Work Practices by Using Augmented Reality. Jusletter IT. 2021:241-7. Available from: https://doi.org/10.38023/e40dcea9-9724-4318-bdca-52ce1cb04e68.

[7] Lachmayer F, Cyras V. Visualization of Legal Informatics. In: Schweighofer E, editor. JURIX2021: The Thirty-fourth Annual Conference, Vilnius, Lithuania; 2021. p. 3-10. Available from: https://doi.org/10.3233/FAIA210310.

[8] Sasdelli D, Steffes B. PROLEG and Normative Diagrams. In: New Frontiers in Artificial Intelligence: JSAI International Symposium on Artificial Intelligence, JSAI-IsAI 2025, Osaka, Japan, May 26–27, 2025, Proceedings. Berlin, Heidelberg: Springer-Verlag; 2025. p. 130–145. Available from: https://doi.org/10.1007/978-981-96-7071-0_9.

[9] Sasdelli D, Steffes B, Herrmann M, Chitashvili M, Wüst C. A Normal Form for Representing Legal Norms and its Visualisation Through Normative Diagrams. Annual International Conference on Digital Government Research. 2025 Jun;26. Available from: https://proceedings.open.tudelft.nl/DGO2025/article/view/1036.

[10] Bobbio N. Teoria generale del diritto. Torino, Italy: G. Giappichelli; 1993.

[11] Kelsen H. Allgemeine Theorie der Normen. Vienna, Austria: Manzsche Verlags- und Universitätsbuchhandlung; 1979.

[12] Steffes B, Sasdelli D. Alternative Darstellungsformen Normativer Diagramme. Jusletter IT. 2025. Available from: https://doi.org/10.38023/51565951-d1ea-4767-877f-639456f882af.

[13] Steffes B, Sasdelli D. A Legal Visualisation Tool Using Normative Diagrams. In: Legal Knowledge and Information Systems - Proceedings of the 37th International Conference on Legal Knowledge and Information Systems – JURIX 2024. IOS Press BV; 2024. p. 395-7. Available from: https://ebooks.iospress.nl/volumearticle/71046.

[14] Steffes B, Sasdelli D. ViNo – A Tool for the Visualisation of Norms. In: Maranhão J, editor. Proceedings of ICAIL2025. ACM; 2025. p. 373-4. Available from: https://doi.org/10.1145/3769126.3769204.

[15] Alchourrón CE. On Law and Logic. Ratio Juris. 1996;9(4):331-48. Available from: https://doi.org/10.1111/j.1467-9337.1996.tb00250.x.

[16] Hart HLA. The Concept of Law. Oxford, UK: Oxford University Press; 1994.

[17] Kelsen H. Reine Rechtslehre. 2nd ed. Vienna, Austria: Franz Deuticke; 1960.

[18] Kelsen H. Reine Rechtslehre. 1st ed. Vienna, Austria: Franz Deuticke; 1934.

[19] Marmor A. The Pure Theory of Law. In: Zalta EN, editor. The Stanford Encyclopedia of Philosophy. Fall 2021 ed. Metaphysics Research Lab, Stanford University; 2021. Available from: https://plato.stanford.edu/archives/fall2021/entries/lawphil-theory.

[20] Governatori G, Rotolo A, Sartor G. Logic and the Law: Philosophical FOundations, Deontics, and Defeasible Reasoning. In: Gabbay D, Horty J, Parent X, van der Meyden R, van der Torre L, editors. Handbook of Deontic Logic and Normative Systems. Vol. 2. London: College Publications; 2021. p. 657-763.

[21] von Wright GH. Deontic Logic. Mind. 1951;60(237):1-15. Available from: http://www.jstor.org/stable/2251395.

[22] Parent X, van der Torre L. Introduction to Deontic Logic and Normative Systems. College Publications; 2018. Available from: https://www.collegepublications.co.uk/TLR/?00001.

[23] Sasdelli D. Können Maschinen Rechtsfälle entscheiden? Hamburg: Felix Meiner Verlag; 2025. Available from: https://doi.org/10.28937/978-3-7873-4900-5.

[24] Sergot MJ, Sadri F, Kowalski RA, Kriwaczek F, Hammond P, Cory HT. The British Nationality Act as a logic program. Commun ACM. 1986 May;29(5):370–386. Available from: https://doi.org/10.1145/5689.5920.

[25] Satoh K. PROLEG: Practical Legal Reasoning System. In: Warren DS, Dahl V, Eiter T, Hermenegildo MV, Kowalski R, Rossi F, editors. Prolog: The Next 50 Years. Cham: Springer Nature Switzerland; 2023. p. 277-83. Available from: https://doi.org/10.1007/978-3-031-35254-6_23.

[26] Sartor G. Legal concepts as inferential nodes and ontological categories. Artificial Intelligence and Law. 2009 Sep;17(3):217-51. Available from: https://doi.org/10.1007/s10506-009-9079-7.

[27] Robaldo L, Batsakis S, Calegari R, Calimeri F, Fujita M, Governatori G, et al. Compliance checking on first-order knowledge with conflicting and compensatory norms: a comparison among currently available technologies. Artificial Intelligence and Law. 2024 Jun;32(2):505-55. Available from: https://doi.org/10.1007/s10506-023-09360-z.

[28] Kant I. Grundlegung zur Metaphysik der Sitten [Kants Werke Bd. 4, Kritik der reinen Vernunft (1. Aufl. 1781), Prolegomena, Grundlegung zur Metaphysik der Sitten, Metaphysische Anfangsgründe der Naturwissenschaft]. De Gruyter; 1968.

[29] Lopes PM. Legal Norms as Hypothetical Imperatives. In: Sellers M, Kirste S, editors. Encyclopedia of the Philosophy of Law and Social Philosophy. Dordrecht: Springer Netherlands; 2017. p. 1-8. Available from: https://doi.org/10.1007/978-94-007-6730-0_219-1.

[30] Maier H. Psychologie des emotionalen Denkens. Tübingen: Mohr & Siebeck; 1908.

[31] Chisholm RM. Contrary-To-Duty Imperatives and Deontic Logic. Analysis. 1963 12;24(2):33-6. Available from: https://doi.org/10.1093/analys/24.2.33.

[32] Hilpinen R, McNamara P. Deontic Logic: A Historical Survey and Introduction. In: Gabbay D, Horty J, Parent X, van der Meyden R, van der Torre L, editors. Handbook of Deontic Logic and Normative Systems. College Publications; 2013. p. 3-136. Available from: https://www.collegepublications.co.uk/downloads/handbooks00001.pdf.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251588*

# That's So FETCH: Fashioning Ensemble Techniques for LLM Classification in Civil Legal Intake and Referral

Quinten STEENHUIS [a,1]
[a] *Suffolk University Law School, Boston, MA*

**Abstract** Each year millions of people seek help for their legal problems by calling a legal aid program hotline, walking into a legal aid office, or using a lawyer referral service. The first step to match them to the right help is to identify the legal problem the applicant is experiencing. Misdirected applicants may miss a deadline, experience physical abuse, lose housing or lose custody of children while waiting to connect to the right legal help. We introduce and evaluate the FETCH classifier for legal issue classification and describe two methods for improving accuracy: a hybrid LLM/ML ensemble classification method, and the automatic generation of follow-up questions to enrich the initial problem narrative. We employ a novel dataset of 419 real-world queries to a nonprofit lawyer referral service. We achieve 97.37% accuracy (hits@2) with inexpensive models, slightly above GPT-5, while reducing cost of guiding users to the right legal resource.

**Keywords.** llm, ensemble classification, legal aid, access to justice, legal intake

## Introduction

Finding the right lawyer can be hard. Even those who know a lawyer may find they practice in the wrong area, while low-income applicants struggle to identify the right aid program and others must rely on friends or referral services. Getting legal triage or referral wrong means that the applicant may lose access to timely legal help and risk liberty, physical abuse or injury, loss of housing, or custody of a child.

One tool for classifying natural language text is a large language model (LLM). Even current state-of-the-art LLMs do not classify with 100% accuracy. "Ensemble" classification, which combines multiple classification methods to produce a single score [7], is a well-known method of improving classification accuracy.

This paper introduces the FETCH (Fast Ensemble Tagging and Classification Helper) classifier. FETCH is a multi-label weighted ensemble voting classifier that combines LLMs, keyword matching, and traditional ML.

[1]Corresponding Author: Quinten Steenhuis, Suffolk University Law School, 120 Tremont St, Boston, Massachusetts, USA; E-mail: qsteenhuis@suffolk.edu. https://orcid.org/0009-0001-0110-064X

This work was produced in collaboration with the Virginia Legal Aid Society and the Oregon State Bar. Virginia Legal Society (VLAS) receives approximately 18,000 calls to human intake workers each year. The Oregon State Bar (OSB) uses trained staff to field approximately 100,000 inquiries to its bar referral service each year.

Ultimately, we show that a combination of three inexpensive LLM models: GPT-5-nano, gemini-2.5-flash, and mistral-small, together with keywords and the traditional ML Spot API [8] may meet or exceed the performance of the current state-of-the-art GPT-5 model at a significant cost and speed savings.

### *0.1. Research questions*

- RQ1: Can small-model ensembles deliver equivalent or superior classification accuracy to frontier models while reducing cost in legal intake?
- RQ2: What types of legal intake cases remain most challenging for ensemble and LLM classifiers, and how can error analysis inform classifier design?

## 1. Prior work

### *1.1. AI for access to justice*

The application of AI to the domain of access to justice is now the topic of a biannual workshop co-located with JURIX and ICAIL [14]. See, e.g., applications to legal information tools [6,25], form-filling and expert systems [19,18,24], and dispute resolution [23,4,5].

### *1.2. Ensemble classification of natural language text*

Ensembles have been used to improve accuracy when classifying text for decades [7,11], and LLMs have extended ensemble techniques to medicine [13] and law [9,12,16].

### *1.3. Legal intake and advice*

[20] applied LLMs to legal intake, specifically threshold eligibility determinations in the legal aid context. The work in this paper builds on and generalizes that prior work.

## 2. The legal problem classification task

An attempted legal referral can end in one of three ways:

- The applicant is matched with a provider who specializes in the kind of problem the applicant is experiencing.
- The intake worker may decide the applicant has a problem that does not have a legal solution, due to inapplicability, statutes of limitation, or jurisdiction.

- The intake worker may decide that the applicant is not experiencing a legal problem at all, but instead may be under a delusion.[2] In these cases, the intake worker may pay extra attention to handle the applicant's problem with sensitivity.

Classification is central to each of these tasks. Applicants may have an idea of the kind of lawyer they need to help with their problem, but that idea may be wrong.[3] Informally, we observed applicants struggle when faced with distinguishing between abstract legal categories, and this is supported by studies such as [26] (showing that only 29% of people experiencing a legal problem recognize it as such) and [10] (emphasizing that self-assessed legal understanding often diverges from actual understanding).

Perfect referral is not always possible in a given geographical area, and applicants may settle for a next-best match in those cases.

## 3. The FETCH classifier

The FETCH classifier is available as a REST-like API, in use by a phone intake system for Virginia Legal Aid Society and the Oregon State Bar's lawyer referral service.

We employ a single prompt across three LLMs.[4] The prompt includes the taxonomy and the applicant's problem description to determine if:

1. The problem can be classified into one or more of 244 taxonomy nodes.
2. More information is needed to accurately classify the applicant's problem.
3. The applicant has not described a legal problem at all.

We experimented with different levels of description in the taxonomy for the prompt but settled on an abbreviated level, with simply the name of each category and its parent category. For example: a characteristic row reads simply "Administrative Law > Utilities".

The results are scored by the number of times each node appears in the results of one of the models. The top two nodes are shown to the applicant.

If the ensemble does not have sufficient confidence in the classification, each LLM model in the ensemble is instructed to generate 3 follow-up question candidates. The prompt was adjusted iteratively to balance generating follow-up questions against making a best effort to classify the applicant's problem.

## 4. Oregon State Bar's referral form

One use of the FETCH classifier for the Oregon State Bar's online referral system (Figure 1). As shown in Fig. 1, applicants interact with the system through a consistent form metaphor, rather than in the form of a chatbot. We hypothesize that the form both reduces effort for the applicants and reduces the "uncanny valley" [22] effect that users experience when talking to an LLM [15].

[2] A classic, commonly experienced example described by staff at the Oregon State Bar was "I have a microchip in my head". The phenomenon of body-control delusions has a well-documented link to paranoid schizophrenia dating to 1919. See, e.g., [21].

[3] For example, a layperson might be asked to decide that their problem getting Social Security Disability benefits falls under the category "Administrative Law" (a relatively obscure term for a layperson).

[4] https://gist.github.com/nonprofittechy/ef36fb8da928f25c60cd0ecb82a80750

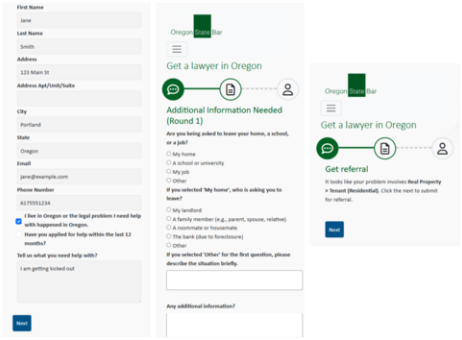



**Figure 1.** Screenshot of the lawyer referral service with integrated FETCH classifier, showing an automatically generated follow-up information form generated for an ambiguous query.

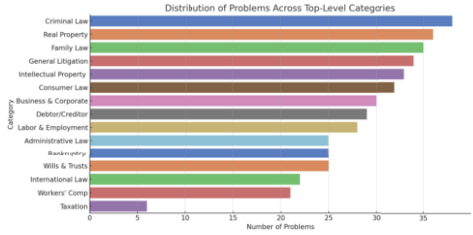



**Figure 2.** Distribution of problem type by top-level category name in the dataset.

Because it generates follow-up questions automatically, the FETCH classifier enables the tool to gather more information without a human having to call the applicant.

Applicants who are screened out are given the option to call a human intake worker at the Oregon State Bar.

## 5. Experimental design

### *5.1. Dataset*

In our experiment, we employed a novel dataset of 419 human-annotated queries made to the Oregon State Bar Referral Service over a period of a few weeks in 2025. The dataset contains a total of 31,716 words, with a mean problem description length of 74 words. Approximately 300 of the queries were annotated to one of the 15 top levels of the legal taxonomy only, while the 119 remaining queries were annotated to a specific "terminal node" in the OSB taxonomy. Annotation was performed by legal professionals at the Oregon State Bar during the regular operation of their referral program. The dataset was anonymized by the bar referral service prior to our analysis.[5]

Queries in the dataset show a diversity in length (Fig. 3) and detail. Spelling errors and idiosyncratic punctuation appear throughout.

[5]A small subset with additional deidentification is available publicly at https://gist.github.com/nonprofittechy/2aa8bfc7141569c7fdd7591564d8e270

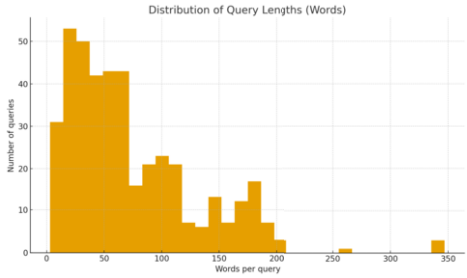



**Figure 3.** Distribution of applicant query lengths in the dataset.

Examples ranged from brief (e.g., Need bankruptcy lawyer) to high-stakes problems, from stalking and domestic violence abuse to emergency housing.[6]

### *5.2. Taxonomy*

Our study uses a 244 node taxonomy maintained by the Oregon State Bar,[7] but national in its application. This represents a superset of alternative issue taxonomies such as the 135-node Legal Matter Specification Standard (commonly referred to as SALI taxonomy) [17] and the smaller civil legal aid-focused LIST Taxonomy.[8]

### *5.3. Classifiers*

Our study encompasses 9 separate classification approaches:

- Keyword matching.
- A term frequency/inverse document frequency classifier.
- GPT-4.1-mini and GPT-4.1-nano.
- Gemini-2.5-flash.
- Mistral-small-3.2-24b-instruct.
- Suffolk Legal Innovation and Technology Lab's traditional ML Spot [8] classifier (build 10).
- GPT-5 and GPT-5-nano, released near the end of our study, are the only "thinking" models in the ensemble.

We deliberately selected a diverse group of providers, models and approaches for the ensemble. The voting method is simple weighted vote, with arbitrary weights initially assigned based on the rough performance of each classifier across a subset of the training dataset. No reweighting was performed after the initial arbitrary starting weights. The classifier was limited to providing no more than the 2 "best" labels based on the weighted ranking.

When available in the response, as for SPOT and for Gemini, via logprobs as suggested by [9], we use the confidence level as an additional weighting factor.[9]

[6] Unlike in [20], we did not observe any instances of model "censorship" or refusal to classify situations involving violence or abuse, a promising indication of the suitability of LLMs for problems in the legal domain.

[7] For the full taxonomy, see https://gist.github.com/nonprofittechy/c8a85248d3835d77193f88d1ff41eda3

[8] https://taxonomy.legal/

[9] Unfortunately, the GPT-5 model series from OpenAI no longer makes logprobs available.

### 5.4. Evaluation method

We evaluated the raw performance of the classifier with the help of the open-source Promptfoo tool against the human-annotated dataset with a hits@2 scoring system. We scored a result as "pass" if the human-annotated label was present in the response, and scored it as passing with a score of 0.5 if at least the top-level node was present in the labels.

## 6. Results

### 6.1. Hits@2 results

Table 1 shows the results as scored using hits@2. The keyword classifier set a baseline at 54%, while TF-IDF fared worst. Notably, the results show that the ensemble method performs best, meeting or slightly exceeding the performance of best-of-class GPT-5 alone[10] and greatly exceeding the average performance of each inexpensive model used in the ensemble classifier.

| Method | Hits@K (top 2 labels) (%) |
|---|---|
| Spot [8] | 59.43 |
| Keyword | 54.18 |
| TF-IDF | 31.03 |
| Gemini 2.5-flash | 87.35 |
| Mistral small | 87.83 |
| GPT-4.1-mini | 81.38 |
| GPT-5-nano | 87.11 |
| GPT-5 | 96.66 |
| Ensemble (GPT-5-nano, Gemini 2.5-flash, Mistral small, Keyword, Spot) | **97.37** |

**Table 1.** Model and ensemble accuracies.

Because the system's interface presents applicants with the top two categories, hits@2 reflects the practical success criterion more directly than top-1 accuracy. The goal is not to force a single label, but to offer two plausible options so that the applicant can choose the most appropriate referral.

### 6.2. Cost efficiency

An important justification for the use of weighted ensemble voting is cost efficiency. As shown in Table 2, the ensemble voting method outperforms the state-of-the-art GPT-5 model on a cost basis by a factor of more than 3.

The latency of the small LLMs is also much lower. For example, GPT-5 has a minimum latency of over 5 seconds [1] compared to GPT-5-nano's minimum latency of 2.2s [2].

[10] While suggestive, the .71 margin between GPT-5 and the ensemble is likely to be statistically insignificant.

| Model | Cached ($/1k) | Uncached ($/1k) | Output ($/1k) | Avg. cost ($/1k req) | Est. annual cost 100k ($) |
|---|---|---|---|---|---|
| mistral-small-3.2-24b-instruct | 0.026 | 0.008 | 0.030 | 0.064 | 6.40 |
| gemini-2.5-flash | 0.039 | 0.048 | 0.750 | 0.837 | 83.75 |
| GPT-5 | 0.065 | 0.201 | 3.002 | 3.267 | 326.71 |
| GPT-5-nano | 0.003 | 0.008 | 0.120 | 0.131 | 13.07 |
| GPT-4.1-nano | 0.013 | 0.016 | 0.120 | 0.149 | 14.91 |
| GPT-4.1-mini | 0.052 | 0.064 | 0.480 | 0.596 | 59.64 |

**Table 2.** Cost model assumes a typical query of **519.5 cached input**, **160.5 uncached input**, and **300.2 output** tokens. Provider assumptions: OpenAI cached input is priced at **10% of the input rate**; Gemini 2.5 Flash cached input is **$0.075 per 1M tokens**; OpenRouter `mistral-small-3.2-24b-instruct` treats cached input at the **same price as input**. Note that the Spot API is free for nonprofits.

### *6.3. Errors*

We closely examined the 11 errors of the ensemble classification method. As in [20], we discovered probable human annotation errors through the LLM classification, totaling four such errors. We also discovered two duplicate entries (omitted in Table 1) in the human annotated dataset which may overweight the LLM classification errors.[11] The full error results are summarized in Table 1. [12]

For the two errors we classified as clear "LLM error," the LLM response was a form of "general litigation" instead of a more specific category. FETCH can optionally generate follow-up questions, and in four of the error results it did generate apparently relevant questions. We therefore examined how these automatically generated questions perform in practice.

### *6.4. Follow-up questions for ambiguous queries*

Many real-world applicant queries are short, vague, or ambiguous. In other cases, while the applicant may describe a situation with multiple legal issues, they may only be able to find an attorney to help with one. Automatic follow-up questions can help clarify the applicant's intent and make legal distinctions that the applicant may not be aware of.

The FETCH classifier generates and selects up to three follow-up questions when confidence is low.[13] The applicant's response is then submitted for an enriched attempt at classification.

To qualitatively test the value of this follow-up question mechanism, we selected 3 candidate questions from the dataset and treated them as follows: one question had details artificially removed. A second question was unchanged, but had been classified by OSB as someone without a legal problem. A third question was misclassified by the LLM. Then we reviewed the output of FETCH for relevance and helpfulness of the follow-up questions to reach an appropriate classification. The results are analyzed in Table 4.

[11] We leave these annotation errors in our tops@2 score for consistency.

[12] It is suggestive, although not conclusive, that removing the human annotation and ambiguous classification errors from the results increases the performance of the ensemble to 99.52%.

[13] https://gist.github.com/nonprofittechy/55ab48c71906930ae5f42dede80df379

| Short summary | Human annotation | LLM | Error classification |
|---|---|---|---|
| Roofing replacement contract; contractor obligations/next steps | Realty > Construction | General, Consumer | LLM error |
| Neighbor harassment/stalking (incl. bear spray); safety concerns | Realty > Neighbor Disputes | General > Neighbor Disputes/, General > Personal Injury | Synonymous |
| Dispute over dog ownership/possession with ex-partner; access withheld | Consumer | General > Animal , Family > Divorce | Ambiguous |
| Wedding planner using event photos/video for promotion; cease-and-desist requested | Consumer | General > Torts/Privacy, Intellectual Property > Trademark | Ambiguous |
| Lost trial for mobile-device while driving; seeks appeal | Consumer | Criminal > Appeals, Criminal > Traffic Offenses | Human error |
| Wants help filing a small-claims case | Criminal | Consumer > Small Claims Advice, General | Human error |
| Car accident; likely uninsured; facing costs; asks about payment plans | Debtor/Creditor | General > Personal Injury, General > Property Damage | LLM error |
| Pandemic unemployment benefits issue; denial/overpayment and appeal options | Debtor/Creditor | Administrative > Unemployment, Labor & Employment > General | Human error |
| Physician Assistant practice without supervising MD/DO; licensing/supervision compliance | Labor & Employment | Administrative > Professional Licensing, Business > General | Human error |

**Table 3.** Detailed error results.

## 7. Discussion

### *7.1. RQ1: Can small-model ensembles deliver equivalent or superior classification accuracy to frontier models while reducing cost in legal intake?*

We have shown that small-model ensembles substantially improve on the performance of older machine learning and keyword-matching approaches to classifying legal problems, and meet or slightly improve on the performance of frontier LLM models. What's more, they do so with lower latency and at a significantly lower cost, better than $\frac{1}{3}$ the cost of using a frontier model for the same task. There was no evidence of bias in the distribution of errors and safety concerns revealed by the errors appear to be minimal.

Importantly, we express the taxonomy to the LLM simply as a list of the taxonomy names, without any additional explanation. This offers promise that applying the technique to future classification tasks may not require extensive prompt engineering, and that it would expand beyond the context of Virginia and Oregon.

Our approach shows promise to create a cost-effective way to deliver high quality referral or eligibility determination information to applicants for legal help.

| Query | Generated Follow-up Questions | Analysis |
|---|---|---|
| I need help with my baby. There's a domestic situation | 1. Is this about the baby's parents or guardians who are or may be divorcing or separating?<br>2. Is there safety risk now for you or your baby, or are you worried about safety? (Options: Yes, there is immediate danger, There is some risk but not immediate, No risk currently, Not sure)<br>3. Do you need help with child custody or visitation for your baby? (Options: Yes, No, Not sure) | (Context removed). The generated questions identify the two potential legal issues: divorce and restraining order, and a correct answer would match the original classification. |
| I have degenerative arthritis in my ankle and have to wear a brace, making it hard to walk unless I use a walker | 1. Are you seeking financial assistance or benefits due to your condition?<br>2. Has your condition affected your ability to work or keep a job?<br>3. Are you experiencing any problems with accessibility in public places or at your home because of your difficulty walking? | (full question, no attorney assigned by OSB). The clarifying questions are relevant and identify potential legal issues, such as denial of benefits or Americans with Disabilities Act compliance. |
| There was a car accident [...] I saw a letter, explaining I owe $10,000 due to the car crash, [...] if any payment plans could be done and I need a lawyer for legal advice. | 1. What is the current status of this case? (Options: Lawsuit filed or court filing, Demand letter or collection letter only, Not sure/unknown)<br>2. Were you at fault in the car accident? (Options: Yes, No, Unsure)<br>3. Was there a police report about the accident? | (full question, misclassified by LLM as dealing with the car crash) The clarifying questions do get us closer to the human annotator's decision that despite originating with a car accident, at core this is a debtor/creditor problem. |

**Table 4.** Examples of ambiguous applicant queries and clarifying follow-up questions generated by FETCH.

### *7.2. RQ2: What types of legal intake cases remain most challenging for ensemble and LLM classifiers, and how can error analysis inform classifier design?*

We did not observe any patterns in the errors in our ensemble classifier's results. They did not appear to be biased towards one type of legal problem over another, and similar categories were not consistently confused by the ensemble. Importantly, this tells us that enriching the category descriptions or adding additional clarifying instructions in the prompt is not likely to significantly improve classification performance.

Overall, annotation errors and ambiguities suggest human baseline performance may match the ensemble. Initial qualitative results suggest the follow-up questions would further improve the performance of the system in real-world usage.

Overall, the errors made by the LLM appear to be remarkably few (approximately 2%), and consistent with the kind of challenges humans may face in similar circumstances.

### *7.3. Limitations and risks*

The results of the study are strong. However, there are some limitations:

- While a relatively large dataset in this domain, 419 queries may not adequately cover the 244 nodes in the full taxonomy. Our findings are strongest when considering the 15 top-level categories.
- During our study, OSB was only able to collect 19 "no legal problem" scenarios, not enough for a robust analysis, and did not provide clear criteria for declining to assign an attorney. Therefore we did not fully test the "no legal problem" condition, especially delusional queries, which will require future sensitivity testing.
- We did not measure whether faster automated referral improved real outcomes.
- Bias is an important risk to consider when using LLMs [3], and biased classification may lead to unfair allocations of scarce low-cost and free legal help resources. These concerns are stronger when the LLM has a higher error rate and particularly when the errors show a pattern that is uneven across the distribution of applicants and problem types, which we did not observe.

## 8. Conclusion and future work

The ensemble method shows promise to improve performance on a high-stakes classification task. Our results show that a weighted combination of inexpensive small language models can slightly outperform the current state-of-the-art model, GPT-5, at approximately $\frac{1}{3}$ of the cost.

Although the error rate is small in our study, future work could further narrow the gap. One promising technique is a routing method, where a frontier model like GPT-5 is asked to classify only the "hardest" cases. We additionally hope to run a broader, quantitative study of the value of follow-up answers in improving classification. Gathering information about the baseline human error rate would add additional value.

## Acknowledgments

The author thanks Virginia Legal Aid Society and Oregon State Bar for funding and data access and David Colarusso, Jack Adamson, and Jim Graszer for their contributions and support of this project.

## References

[1] Gpt-5 api, providers, stats. `https://openrouter.ai/openai/gpt-5`, 2025. GPT-5 is OpenAI's most advanced model, offering major improvements in reasoning, code quality, and user experience. It is optimized for complex tasks that require step-by-step reasoning, instruction following, and accuracy in high-stakes use cases. Run GPT-5 with API.

[2] Gpt-5 nano api, providers, stats. `https://openrouter.ai/openai/gpt-5-nano`, 2025. GPT-5-Nano is the smallest and fastest variant in the GPT-5 system, optimized for developer tools, rapid interactions, and ultra-low latency environments. While limited in reasoning depth compared to its larger counterparts, it retains key instruction-following and safety features. Run GPT-5 Nano with API.

[3] Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2021), FAccT '21, Association for Computing Machinery, p. 610–623.

[4] BICKEL, E. A., VAN DIJK, M. A., AND GIEBELS, E. Online legal advice and conflict support: A dutch experience. *Report, University of Twente* (2015).

[5] BRANTING, K., MCLEOD, S., HOWELL, S., WEISS, B., PROFITT, B., TANNER, J., GROSS, I., AND SHIN, D. A computational model of facilitation in online dispute resolution. *Artificial Intelligence and Law 31*, 3 (2022), 465–490.

[6] BRANTING, L. K. Advisory systems for pro se litigants. In *Proceedings of the 8th international conference on Artificial intelligence and law* (2001), pp. 139–146.

[7] BREIMAN, L. Bagging predictors. *Machine Learning 24*, 2 (Aug. 1996), 123–140.

[8] COLARUSSO, D. Machine-assisted issue spotting for self-represented litigant portals: An introduction to spot—a multi-label text classifier. Tech. rep., Suffolk University Law School, May 2022.

[9] FARR, D., MANZONELLI, N., CRUICKSHANK, I., STARBIRD, K., AND WEST, J. Llm chain ensembles for scalable and accurate data annotation. In *2024 IEEE International Conference on Big Data (BigData)* (2024), pp. 2110–2118.

[10] OECD, AND INITIATIVE, O. S. J. Legal needs surveys and access to justice. Tech. rep., Organisation for Economic Co-operation and Development, Paris, 2019.

[11] OPITZ, D., AND MACLIN, R. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research* (1999).

[12] PRIYADARSHINI, R., ET AL. Ledocl: A semantic model for legal documents classification using ensemble methods. *Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12*, 9 (April 2021), 1899–1908.

[13] QIU, J., GUO, D., PAPINI, N., PEACE, N., LEVINSON, C. A., AND HENRY, T. R. Ensemble of Large Language Models for Curated Labeling and Rating of Free-text Data, 2025. _eprint: 2501.08413.

[14] QUINTEN STEENHUIS, MARGARET HAGAN, MARC LAURITSEN, HANNES WESTERMANN, JAROMÍR SAVELKA. AI for Access to Justice (AI4A2J) Workshop at icail 2025. `https://ai4a2j.org`, 2025. Full-day, hybrid workshop on AI tools, datasets, and approaches to expand access to justice; held on June 20, 2025 at ICAIL 2025 (Northwestern University, Chicago, IL).

[15] RADIVOJEVIC, K., CHOU, M., BADILLO-URQUIOLA, K., AND BRENNER, P. Human Perception of LLM-generated Text Content in Social Media Environments, Sept. 2024. arXiv:2409.06653 [cs].

[16] SAKAI, H., AND LAM, S. S. Quad-llm-mltc: Large language models ensemble learning for healthcare text multi-label classification, 2025.

[17] SALI ALLIANCE. Legal matter specification standard (lmss). `https://www.sali.org`, 2025. Taxonomy of legal-matter classification released by the SALI (Standards Advancement for the Legal Industry) Alliance.

[18] STEENHUIS, Q. Ai and tools for expanding access to justice. In *The Cambridge Handbook of AI in Civil Dispute Resolution*. Cambridge University Press, forthcoming 2024, p. 17. Available at SSRN: https://ssrn.com/abstract=4876633 or http://dx.doi.org/10.2139/ssrn.4876633.

[19] STEENHUIS, Q., AND COLARUSSO, D. Digital Curb Cuts: Towards an Open Forms Ecosystem. *Akron Law Review 54*, 4 (2021), 2.

[20] STEENHUIS, Q., AND WESTERMANN, H. Getting in the door: Streamlining intake in civil legal services with large language models. In *Legal Knowledge and Information Systems — JURIX 2024: The 37th International Conference* (Dec. 2024), J. Savelka, J. Harasta, T. Novotna, and J. Misek, Eds., vol. 395 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, pp. 155–167.

[21] TAUSK, V. On the origin of the "influencing machine" in schizophrenia. *The Psychoanalytic Quarterly 2* (1933), 519–556. Originally published in German in 1919 in the Internationale Zeitschrift für Ärztliche Psychoanalyse.

[22] WANG, S., LILIENFELD, S. O., AND ROCHAT, P. The Uncanny Valley: Existence and Explanations. *Review of General Psychology 19*, 4 (Dec. 2015), 393–407.

[23] WESTERMANN, H. *Using artificial intelligence to increase access to justice*. PhD thesis, Université de Montréal, Oct. 2023. Accepted: 2023-12-08T19:46:03Z.

[24] WESTERMANN, H. Dallma: Semi-structured legal reasoning and drafting with large language models. In *2nd Workshop on Generative AI and Law, co-located with the International Conference on Machine Learning (ICML)* (Vienna, Austria, 2024).

[25] WESTERMANN, H., AND BENYEKHLEF, K. JusticeBot: A Methodology for Building Augmented Intelligence Tools for Laypeople to Increase Access to Justice. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law* (New York, NY, USA, Sept. 2023), ICAIL '23, Association for Computing Machinery, pp. 351–360.

[26] WORLD JUSTICE PROJECT. Global Insights on Access to Justice: Findings from the World Justice Project General Population Poll in 101 Countries, 2019.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251589*

# How Anonymous Is Anonymous? A Techno-Legal Exploration

Stephanie VON MALTZAN [a,1], Daniel VOGEL [b], Marc OHM [b,c] and Florian IDELBERGER [a]

[a] *Karlsruhe Institute of Technology and FIZ Karlsruhe*
[b] *University of Bonn and Lamarr Institute*
[c] *Fraunhofer FKIE*

ORCiD ID: Stephanie von Maltzan https://orcid.org/0009-0004-8135-4202, Daniel Vogel https://orcid.org/0009-0004-1717-1524, Marc Ohm https://orcid.org/0000-0002-2913-5270, Florian Idelberger https://orcid.org/0000-0002-9253-1524

**Abstract.** This paper proposes a pragmatic exploration to facilitate the categorisation of personal data as anonymous, quasi-anonymous, or pseudonymous, emphasising contextualised threat modelling and proportionality over binary thresholds. Using an integrated legal analysis and system-level threat model, we map legal criteria to the design features and assess whether a privacy-preserving system like DROPS can credibly achieve anonymisation under the GDPR. This allows us to evaluate the discrepancy between the technical realities of maximising anonymisation techniques and the requirements for anonymisation stipulated by the EU data protection law corpus. The distinguishing feature of this paper is its grounding of the legal analysis in the technical architecture, thereby bridging the gap between abstract regulation and system-level design. This demonstration has the potential to serve as a model for enhancing data protection measures, particularly for entities that handle high-risk or otherwise sensitive data and for regulators to issue new concrete guidance on anonymisation.

**Keywords.** Anonymisation, Quasi-Anonymisation, Pseudonymisation, Hashing, Privacy, GDPR, PII data

## 1. Introduction

Amid a rising tide of data breaches and increasingly sophisticated cyberattacks, determining whether personal data has been genuinely anonymised under the General Data Protection Regulation (GDPR) has become a complex and pressing issue. The GDPR does not provide an explicit definition of anonymous data. Instead, it can be inferred from a combined interpretation of Article 4(1) GDPR, which defines personal data, and Recital 26 GDPR, which explains how to assess identifiability. However, the threshold between anonymisation and pseudonymisation remains unclear and debated. Data protection authorities, courts, and legal scholars across Europe have adopted varying interpretations – from

[1]Corresponding Author: Stephanie von Maltzan, stephanie.maltzan@kit.edu

relative to absolute approaches and from strict to functional anonymisation [21,35,26,10] —- and legal and technical experts often view anonymisation differently. In light of this, many organisations either incorrectly assess the effectiveness of their anonymisation processes or struggle to comply. Against this backdrop of uncertainty, this paper offers a techno-legal exploration of anonymisation and its role in data protection regulations. It emphasises the importance of contextual threat modelling and proportionality, rather than adopting an all-or-nothing approach based on binary thresholds such as the absolute and relative approach. Using the DROPS system [14] as a case study, we ask: can a system that employs cutting-edge, privacy-preserving methods credibly claim to anonymise data under the GDPR, or does the data remain personal despite strong safeguards? This is of particular interest given the different views [18,11] on whether pseudonymisation should be considered an effective means of anonymisation. If strict anonymisation cannot be practically achieved, even with robust safeguards in place, should the legal standard be adjusted to take a more pragmatic, risk-based approach? Our analysis underpins legal discussions by examining the technical architecture of DROPS, thus linking abstract regulatory standards to system-level design choices. This integrated approach clarifies GDPR compliance and serves as a model for data protection authorities to issue new, more concrete guidance, as well as enabling the European Court of Justice (CJEU) to consider illegal means in its assessments. We argue that absolute and relative binary categories no longer provide an adequate framework for addressing identifiability under the GDPR. Although we agree with the recent SRB judgement [11] that pseudonymous data can be anonymous in certain instances, we also recognise that illegal means to obtain additional information should be considered. Ultimately, our aim is to foster a more practical understanding of anonymisation through this example.

The paper proceeds as follows: Section 2 reviews work relating to the intersection of anonymisation in law and technology. Section 3 describes the DROPS system architecture and design choices, while the subsequent section assesses DROPS from legal and technical perspectives, applying GDPR and WP29 anonymisation criteria in light of a threat model.

## 2. Related Work

Prior research [28,38,27,39,16,12,33,6,34] has addressed specific aspects of the problem, such as legal critiques of anonymisation under the GDPR [4,29,32,37,31], technical limitations of anonymisation techniques [6,15,22,24], and highlighted key misunderstandings[2]. Achieving meaningful anonymisation in practice remains a significant challenge [25,36, 37] that straddles law and computer science. The most influential regulatory guidance remains the Article 29 Working Party's (WP29) Opinion 05/2014 [3], which introduced a three-part test for anonymisation in conjunction with the '"means" test.[36] In its latest opinions (Opinion 28/2024 [17] and Opinion 01/2025 [18]), the European Data Protection Board (EDPB) has further clarified the '"means" test. Case law of the CJEU – notably the Breyer [7], IAB [10], Scania [8], Olaf [9] and SRB [11] decisions, — has added nuance (and also uncertainty) by interpreting identifiability based on access to additional information, as well as considering whether re-identification by third parties is legally feasible[40]. The interpretative nature of the provision, combined with inconsistent guidelines and case law, has not only led to divergent views among legal scholars but also among regulators. For example, Spain's [1], France's [5] and Italy's [23] data protection authorities tend

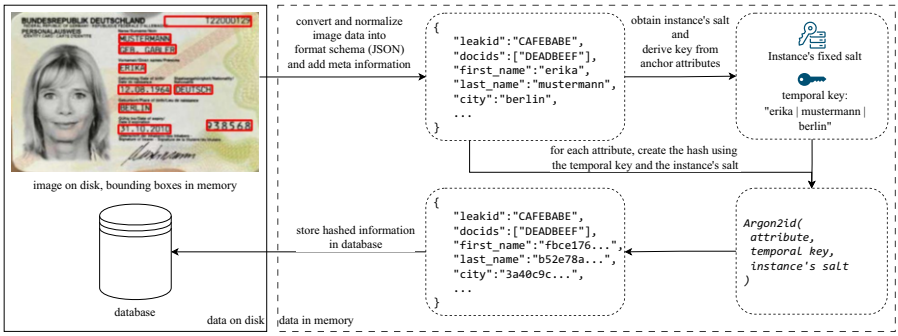



Figure 1.: Flow diagram of data processing within the system.

toward a strict standard, whereas Ireland's [13] has been more lenient in some instances. Such fragmentation underscores the need for a harmonised approach grounded in both sound legal reasoning and practical feasibility, as proposed in this paper. To date, existing research has relied solely on regulatory guidelines, case law or a technical approach, generally accepting the unresolved status quo without conducting an in-depth techno-legal analysis based on a concrete threat model. A key challenge is to demonstrate at what point pseudonymised data might be considered effectively anonymised, and where its ragged edges lie. This paper builds on existing literature by synthesising the perspectives of law and technology and seeks to fill the gap by evaluating the DROPS system against legal criteria using a thread-scenario-based assessment [20], thereby illustrating how and when the GDPR's anonymisation threshold might be reached (or missed) in practice.

## 3. Architecture

The DROPS system[14] is designed to enable comparisons of an individual's data with records of known data breaches (leaks) without exposing any party's raw data to the other. The system accomplishes this through a pipeline of cryptographic hashing and secure comparison protocols, described below and illustrated in Figure 1.

Each personal data value observable in raw identity data is separately hashed using a combination of a system-wide randomly generated secret salt and an anchor, creating unique hashes that do not inherently reveal their link to an anchor to outsiders. The anchor is a value derived from a fixed set of attributes (three in total) for each individual (for instance, a combination of name and address), allowing structured pseudonymisation serving as a identity-unique temporal key. All personal data from a leak dataset is hashed using the memory-hard Argon2id algorithm with the system's secret salt, as recommended by OWASP [30], then truncated. Argon2id's CPU- and memory-intensive nature adds a significant barrier to brute-force or dictionary attacks on the hashes. Importantly, after the hashes are computed, the leak dataset, the original plaintext data (preimage) and the anchor values are immediately discarded. The DROPS database stores only the salted hashes of attributes, each tagged with an internal document ID (doc_id) and a leak ID indicating from which leak the record stems. Neither the anchor nor the preimage is stored in the database, and neither is needed for it to operate. Truncation limits the information an attacker can gather from any single hash value, limiting the effect of precomputed

tables, while only posing negligible collision risk. In fact, truncated Argon2id hashes still enable almost 100 percent accurate matching between datasets through cryptographic protocols like Private Set Intersection (PSI), allowing DROPS and its users to achieve near-perfect intersection accuracy.

DROPS employs a form of PSI to enable secure matching of records between a querying party and the leak database without revealing sensitive data to either side. To initiate a query, an authorised user must use the DROPS client app, which handles the following process client side given correct input: Give anchor attribute values and an ordered list of identity attributes and these attribute's values. Hash each value list entry using anchor as key and the system's salt. Hash the concatenation of these hashes again using the system's salt, then truncate the resulting hash to a pre-negotiated length. Query with truncated hash and ordered list of identity attributes. Upon receiving the query, DROPS compares truncated hashes of attribute sets that share doc_ids. A match occurs only if the full set of queried attribute hashes exists within the DROPS database, assuming no collisions. Throughout this PSI-based protocol, no raw personal data is ever exchanged. The querying party only learns whether the queried attribute hashes exist within the DROPS database. Conversely, the DROPS server learns nothing about the query other than the fact that a query was made – it cannot derive which person the query pertains to.

Through these measures, DROPS establishes a highly secure and privacy-preserving environment. It stores only pseudonymised personal data, requires prior knowledge of key attributes to generate queries, and employs cryptographic protocols so that information disclosure is minimal and controlled. Subsequently, we evaluate how these technical choices translate to GDPR compliance in terms of identifiability of individuals in the data.

## 4. Techno-legal assessment

Under the GDPR, data are considered anonymous if individuals are "not or no longer identifiable" taking into account "all the means reasonably likely to be used" for re-identification. WP29's Opinion [3] established a standard three-criterion test: after anonymisation, it should be impossible to (1) single out an individual in the dataset, (2) link records relating to the same individual (within the dataset or between datasets) relating to the same individual, and (3) infer any additional information about an individual from the data. If any of these risks remain, the dataset cannot be regarded as fully anonymised. Instead, the controller must evaluate the residual risk of identification and determine if it is acceptably low in practice [3,17,20]. Importantly, the GDPR does not demand zero risk — an impossible standard — but any risk of re-identification must be extremely remote. Data can be considered effectively anonymised if re-identifying a person would require a disproportionate effort. This assessment requires consideration of all the means that could reasonably likely to be used by the controller or a third party to identify individuals, and the determination of those means should be based on objective factors such as the cost of identification, time required and technologies available at the time of processing. Anonymisation procedures and their evolution must therefore be continuously reviewed and evaluated [2]. This test requires a comprehensive and contextual risk assessment that goes beyond simply assessing the technical possibility of re-identification and takes into account the practical feasibility and likelihood of such attempts. Consequently, this raises the question of whether there exists additional information, either currently available

or that may become available in the future, that could be utilised in conjunction with the existing data to identify the individual in question effectively. Controllers should first focus on the concrete means (here: original leak dataset, anchor, salt) that would be required to reverse the anonymisation technique, particularly concerning the costs and know-how required to implement these means and the assessment of their likelihood and severity. DROPS must, thus, balance its anonymisation efforts and costs (both in terms of time and resources required) against, for example, the increasingly low-cost availability of technical means to identify individuals in datasets, the public availability of other datasets and the leakage of secret keys, as well as the costs of brute-force attacks. EDPB stresses that risk assessment should account for all actors, including malicious attempts – not only the "legitimate" data recipients foreseen by the controller [18]. This broad view contrasts with the CJEU interpretations [7,8,10,19], which to date have not been particularly explicit about the scope of relevant other persons beyond data recipients concerning the application of the means test. Essentially, courts focus on the intended or anticipated data flows and apparently consider only legal means as applicable. However, this standpoint may be excessively restrictive insofar as it poses a considerable real threat to privacy, especially in a world of frequent data breaches. Following the EDBP's [18] approach, when determining whether data is truly anonymised, it is necessary to consider not only honest actors but also individuals who might deliberately attempt to circumvent de-identification. Our assessment of DROPS follows this approach, examining identifiability from multiple threat angles.

### *4.1. Singling out*

Singling out is "the ability to locate or isolate an individual's record in the dataset" [10]. Even without a person's name, if one can pinpoint a unique record (or set of attributes) that corresponds to the same individual, that individual is singled out.[3] **Access to the DROPS database:** All personal records are individually stored as Argon2id anchor-based hashes along a corresponding `doc_id` (and `leak_id`) to group all attribute-hashes from the same document source. Each hashed identity record corresponds to one real person's attribute (e.g. name) from the leaks. Because DROPS uses deterministic hashing with a fixed salt and anchor per person, each individual's attribute data yields a unique pseudonym (per `doc_id` a set of pseudonymised values) that acts as an identifier for that person within the system. This means that, even though the actual identifiers are hashed, an attacker with access to the DROPS database can use them to distinguish one person's record(s) from another's because each record is stored alongside its `doc_id`. Attackers do not need to know the person's name or the details behind the records to distinguish them. WP29 explicitly noted this pitfall [3] The structure of DROPS inherently allows for singling out individual records in the dataset.

**Through authorised use (a client with query access):** If the user's query finds a match in DROPS, the user has effectively singled out that individual as present in the leak. This is an intended feature outcome, not a bug – but it does highlight that an individual can be singled out (identified as a breach victim) by combining the DROPS processing with the additional knowledge held by the contracting organisation (the individual's identity). From an identifiability perspective, the system deliberately allows known individuals to be singled out (alerted), albeit only by someone who already knew their identity. Therefore, DROPS does not eliminate the risk of singling out, but rather limits who can singling

out records. Notably, the authorised user already knows the identity, DROPS did not reveal a new identity, but it did confirm the existence of that person's data. From a strict legal perspective (although the ability to do anything useful with a singled-out record is severely constrained), DROPS does not eliminate the risk of singling out.

### *4.2. Linkability*

Linkability refers to the ability to link two or more records concerning the same individual within the same dataset or across different datasets [3]. For data to be anonymised, it should not be possible to correlate separate data points and conclude that they are about the same data subject.

**Access to the DROPS database:** The DROPS system's functionality relies on linkability by design. The anchor-based hashing scheme links all attributes of a record under a common key (the anchor, although it is not stored). Internally, this is reflected in the `doc_id` grouping in the database. This internal linking is necessary for the system to function – it allows the PSI query to check for a matching set of attributes. If the same individual's information appears in multiple leaks imported into DROPS, the system will, as long as the personal and anchor information is consistent, assign them the same pseudonym (hash) each time. These hashes will then be assigned multiple `leak_ids`. An insider who inspects the database could notice that two different `doc_id` groups have identical hash values, suggesting that they belong to the same person. This means that DROPS itself can inherently link records across different leaks that pertain to the same individual. Even though the person's name is not stored, the pattern of recurrence is visible. This internal linkability is necessary for the system's purpose. However, this means the second WP29 criterion is not met – the system intentionally preserves the ability to link records referring to the same individual.

**Through authorised use (a client with query access):** Linkability is the core functionality provided to the querying party. When an organisation queries "Jane Doe, Main Street, jd@example.com" (hashed) and DROPS finds a match, it is linking that query record to a leak record. The PSI protocol outputs the intersection, i.e., the links between the set of query hashes and the set of leak hashes. In this way, the authorised user learns that the record in their database (Jane Doe) corresponds to a record in the DROPS database. This linking is done in a privacy-preserving manner (neither side learns anything beyond the existence of the link), but it is linkable across datasets nonetheless (the querying party's list and the DROPS database). A malicious authorised user might try to exploit this by, for example, querying many different partial identities to see which ones exist in DROPS (a fishing expedition to find out who is in the database). DROPS mitigates such abuse by requiring the full anchor for queries and by monitoring for suspicious query patterns. This makes it difficult to systematically probe for links without already knowing the identity of a target individual.

**External attacker:** What complicates linkage for outsiders is that the anchor is secret, and the salt is system-wide but not public. If an attacker did not know about the `doc_id` link, they might not realise that two different-looking hashes actually belong to the same person. However, DROPS uses pseudonyms for different attributes of the same person within the system using the same anchor – for example, all of Jane Doe's data (e.g. name, address) becomes hashed identities under the same anchor context. Even truncated hashes won't break this link, as the hash input data are still linked through the anchor. As the

system maintains consistent pseudonyms the second criterion is not met. This is in line with WP29, which noted that when the same key is used for an individual across records, linkage is "trivial" [3]. An attacker could link records that belong to the same person either by querying using corresponding known personal data and finding a match or by getting access to the DROPS database and filtering by `doc_id`. Neither of these methods, however, allows the attacker to learn any identifiable information that they doesn't already know.

### *4.3. Inference*

Inference is the ability to deduce the value of an attribute with significant probability using other information [3]. DROPS severely limits what can be inferred by anyone without prior knowledge. The data stored and exchanged is heavily transformed (hashed), so that personal attributes cannot be directly inferred; the hash reveals nothing semantic other than its appointed attribute. The data flows are, furthermore, sufficiently controlled that unauthorised inference by outsiders is impossible (since they see nothing but hashed tokens). The authorised parties (the clients), however, infer something non-trivial: they learn that a particular user's credentials were found in a leak. This is arguably new information about that user. If a malicious client identifies the existence of a person's data in a leak, that client could try to find the leak. In fact, the existence of other entries could even be verified by using the leak to gather information about other persons and query their presence in the DROPS database. In that case, the additional knowledge of the fact that the rest of the leaked content is also part of the DROPS database is likely to be insignificant. DROPS performs well against the inference criterion against outsiders – it is extremely resistant to someone inferring new information about individuals. However, under WP29's strict test, the existence of any plausible inference attack means that the data is not anonymised.

By the strict criteria, DROPS does not fully protect against singling out, linkage and inference, and thus cannot be considered anonymisation by this standard. The CJEU's stance that only legal means must be employed would not alter this.

### *4.4. The "all means reasonably likely to be used" test*

While the analysis above finds that DROPS does not satisfy the WP29's strict criteria, the WP216 opinion also states that controllers should carry out an evaluation of the identification risks in order to determine whether the residual risk is acceptable; in other words, whether the anonymisation process is sufficiently robust [3]. For the purposes of this paper, the following simplified analysis considers several attacker models and evaluates the effort, means and likelihood of success for each. This analysis brings us to the concept of quasi-anonymisation, which we introduce to describe data that is not strictly anonymous in law, but is sufficiently de-identified under a contextual, proportionate risk analysis.

**External Attacker: No Access to Salt or Anchor** Given that an attacker gains access to the DROPS database and has access to all hashes associated with their `doc_id` and `leak_id`. An external attacker with sole access to the DROPS database has no direct means of re-identifying individuals using the DROPS database alone. If the attacker had access to various leaked data, some of which were used for the DROPS database,

the attacker could try to create hashes based on the known leaked data. These hashes could be used to find matching hashes in the database, thereby de-pseudonymising the entry. Without knowledge of the salt, re-identification would require the attacker to guess the salt, which would feasibly require an exhaustive search through all possible salts. Since Argon2id is set up to limit the amount of hashes per second that can be calculated, finding the correct salt is a disproportionate effort, as shown by the following calculation: Assuming a DROPS system uses a salt with a length of 32 Byte or 256 Bit and that an individual record can be hashed with Argon2id in 0.01 seconds, meaning 100 hashes per second. An exhaustive search for all values of the salt would require an average of 4.4e68 years using one processor. As the database stores only truncated hashes, the attacker would also have to verify a matching salt for one record with at least a second record, adding an additional step that is somewhat negligible considering the hashing cost. Given the computational intensity of Argon2id and the entropy added by both the salt and the anchor, brute-force reversal is not practically feasible. This would require expensive computational resources. Recital 26 GDPR explicitly directs consideration of the cost and time required, taking into account current technology. Here, an external attacker would need astronomical resources to guess each unknown salt/anchor and recompute Argon2id hashes. Under the Recital 26 test, the means of such an attacker are not reasonably likely – the effort and cost of re-identifying identities are disproportionate, and the identification thus fails the threshold of objective identifiability. In short, to an external attacker without access to anchors and salt (additional information), the DROPS data – although technically pseudonymised — can be deemed, for all practical purposes, anonymous under current technological constraints: without the secret keys, the data cannot be linked to individuals by any feasible method. In terms of considering pseudonymous data as anonymous, this is consistent with the recent SRB judgment [11], but inconsistent with EDPB guidance [18]. We argue that the remaining risk appears to be acceptable. Recital 26, however, demands a holistic assessment, meaning that DROPS must also consider actors who might obtain the additional information.

**Malicious Insider (Access to database, salt):** If an attacker has access to both the DROPS database and the salt, either as a malicious insider or through a salt leak, the means of identification are much more accessible. This attacker would be enabled to generate hashes for identity records for which the anchor attributes are known. It would be possible to use acquired identity data to look for matching hashes, thereby inferring which identity records are stored within the DROPS database. Notably, the attacker does not learn any identity data that is not already known. Nevertheless, it is possible to attribute the data to a specific individual by leveraging additional information, such as a secret salt and the individual's identity details. From a legal perspective, the presence of an individual (the insider) who possesses or obtained a key to unlock the pseudonymised data indicates that the data cannot be regarded as fully anonymised, unless the effort expended is disproportionate. This illustrates that DROPS data is only as anonymous as its secret salt remains secret, as long as the salt and anchor knowledge are confined to the controller's internal system. However, if an insider reveals that secret, the hashes become identifiers, which necessitates treating such data as pseudonymised, with robust technical and organisational safeguards in place, though not to the extent of irreversible anonymisation. In summary, DROPS has been designed to make re-identification of the hash records a complex, but not disproportionately impossible, undertaking in the event of salt leakage. Beyond known identities, guessing unknown identity records involves

the complexity of matching attributes and the correct anchor, which may be subject to hash collisions, taking hash truncation into account. It is arguably less disproportionate to find leaks presented to DROPS that researchers have found on the internet than to guess identity records. However, this would be impossible for leaks shared only by a whistleblower with DROPS and no third party. Without additional information, the complexity of making valid guesses increases significantly due to anchor-hash binding and hash truncation and would require disproportionate brute-force attacks or contextual guessing. Ultimately, without access to the original leak or external auxiliary data, an attacker cannot meaningfully narrow the input space. From a legal perspective, this constitutes a disproportionate effort, rendering the data effectively anonymous. However, if it could be possible to access the original leak, the threshold of disproportionate effort may not be reached. In such cases, the concept of quasi-anonymisation is particularly relevant and is used to describe data that is not strictly anonymous in a legal sense, but is considered sufficiently de-identified under a contextual and proportional risk assessment – especially when re-identification is close to the threshold of disproportionate effort. Even without full anonymisation and with a hashing technique that essentially amounts to pseudonymisation, the contextual risk of re-identification here is deemed low enough to nearly reach the threshold of disproportionate effort. As a result, the data can be classified as outside the scope of the GDPR's definition of personal data. Nevertheless, this status must undergo regular reassessment to ensure its ongoing relevance amid technological advancements and shifting threat landscapes, as well as to incorporate new information. Argon2id is currently the most robust option; however, this could change in the future. Anonymisation should not be seen as a one-off technical achievement but rather as a dynamic, context-sensitive process. Taking the CJEU's perspective into account, it would also be anonymous, since illegal means are not considered. However, according to the EDPB, pseudonymous techniques cannot be attributed to anonymous data.

**Malicious Insider (Anchor, Salt, API key):** Having access to a correct API key would allow an attacker to impersonate a DROPS client and create queries to attack the DROPS system either through database probing, DDoS or similar attacks, which we won't discuss as they are outside the scope of this paper. Probing the DROPS database using a known API key, salt, and prior knowledge of identity anchor information could be used to attempt to learn about the existence of specific attribute records of previously unknown identities. The DROPS database could be used as an oracle to verify a guess, provided the attribute is present in the database. When combined with the salt and anchor knowledge, this creates a feasible attack surface, especially when guessing small attribute spaces. Notwithstanding the implementation of technical and organisational safeguards, re-identification remains a realistic possibility. Consequently, the data remains pseudonymised personal data, albeit not from the perspective of the CJEU.

**Authorised User of DROPS (Querying via Client):** A malicious authorised user who interacts with the system only through the DROPS client would have access to an API key, as well as the DROPS client software. From a GDPR perspective, the limited and cryptographically shielded interaction reduces identifiability for the querying user to a negligible level. Users are constrained to queries for which they possess the anchor information, i.e., prior knowledge of the individual. Consequently, the users are unable to re-identify the individual, as they are already aware of their identity. As re-identification is predicated on pre-existing knowledge, and no further indirect identifiers or structural links are revealed, the DROPS data is (besides the querying hashes) functionally anonymous

for the querying user, and vice versa for DROPS, while the DROPS data is pseudonymous for DROPS. This is in line with the IAB's [10] reasoning that the status of personal data is relative to the information held by each party: one entity's data may be considered personal, while another's may be anonymous.

Incorporating these technical scenarios into legal analysis reveals a pattern: Due to the intended use of DROPS, full anonymisation cannot (which is legally debatable under different approaches) be achieved in all cases – a situation that frequently arises when the data's utility is still necessary. DROPS occupies a middle ground: quasi-anonymous – the data is functionally anonymous given the realistic threat environment and existing safeguards, though not irrevocably anonymised against all attackers, reflecting a pragmatic and more practical balance between utility and privacy. Although residual risks persist, they are effectively mitigated.

Quasi-anonymisation acknowledges that, while full anonymisation may be unattainable in specific contexts, strong technical and organisational (even pseudonymisation techniques) safeguards can significantly reduce the risk of re-identification to an acceptable level. This distinction recognises that anonymisation is not a binary phenomenon, but rather a spectrum. It is proposed that, if quasi-anonymisation is achieved, the GDPR would not apply in advance, but rather steps would be taken to limit the compliance gap and associated risks of non-compliance with the transformation into personal data.

## 5. Conclusion

By applying the "means" test through a simplified attacker cost-and-effort model, we demonstrated the different outcomes of applying the CJEU's and EDPB's guidance, and showed the importance of conducting context-specific risk analyses to determine when data can be considered effectively anonymised. The proposed concept of quasi-anonymisation as a legally relevant category highlights a pragmatic middle ground where data protection can be robust without requiring perfection. The effectiveness of anonymisation will be demonstrated further in future, quantitatively. This paper aims to encourage discussion on widespread practical anonymisation, encourage companies to adopt advanced data protection practices and regulators to reach a consensus. We argue that EDPB should provide concrete guidance that outlines, first, clear criteria for anonymisation and, second, the flexibility to implement quasi-anonymisation measures. If the criteria remains too absolute or unclear, there is a risk that controllers will either misapply it or disregard anonymisation efforts altogether. A more apparent acknowledgement of contextually"'safe" data would incentivise the adoption of strong technical measures by offering some relief under the law when the re-identification risk is minimal. Future work and policy should focus on refining the concept of quasi-anonymisation and developing best practices for achieving it within the GDPR's framework. Importantly, our assessment also sounds a note of caution: methods used in DROPS, while significantly reducing risk, do not entirely eliminate the possibility of re-identification, especially if additional information becomes available. Controllers should avoid any false sense of security. Relying on a single technique is not enough; instead, a defence-in-depth strategy is imperative. Such a holistic approach is essential to maintain effective anonymisation over time, ensuring compliance with the GDPR's demands for data protection by design and by default, and preventing non-compliance in the event of the transformation of data into personal data.

## References


[1] AEPD. Anonymization III: The risk of re-identification, 2025. Accessed August 10, 2025. https://www.aepd.es/en/prensa-y-comunicacion/blog/anonymization-iii-risk-re-identification.

[2] AEPD and EDPS. 10 misunderstandings related to anonymisation. Guidance document, European Data Protection Supervisor and Spanish Data Protection Agency, 04 2021.

[3] Article 29 Data Protection Working Party. Opinion 05/2014 on anonymisation techniques. Technical Report WP216, European Commission, 4 2014. 0829/14/EN.

[4] Andrew Burt, Sophie Stalla-Bourdillon, and Alfred Rossi. A guide to the eu's unclear anonymization standards, 2021. Accessed August 10, 2025. https://iapp.org/news/a/a-guide-to-the-eus-unclear-anonymization-standards.

[5] CNIL. Decision n°san-2023-00 Doctissimo, 2023.

[6] Aloni Cohen and Kobbi Nissim. Towards formalizing the gdpr's notion of singling out. *Proceedings of the National Academy of Sciences 117(15), 8344–8352*, 2020.

[7] Court of Justice of the European Union. Case C-582/14 Breyer, 10 2016.

[8] Court of Justice of the European Union. Case C-319/22 Scania, 11 2023.

[9] Court of Justice of the European Union. Case C-479/22 Olaf, 09 2024.

[10] Court of Justice of the European Union. Case C-604/22 IAB, 3 2024.

[11] Court of Justice of the European Union. Case C-413/23p SRB, 09 2025.

[12] Lorenzo Dalla Corte. coping personal data: Towards a nuanced interpretaton of the material scope of eu data protecton law. *European Journal of Law and Technology, Vol 10, Issue 1*, 2019.

[13] Data Protection Commission Ireland. Guidance on anonymisation and pseudonymisation, 2019.

[14] DROPS Project. DROPS Research Project, 2025. Accessed August 10, 2025. https://itsec.cs.uni-bonn.de/drops/en/.

[15] Khaled El Emam and Cecilia Alvarez. A critical appraisal of the article 29 working party opinion 05/2014 on data anonymization techniques. *International Data Privacy Law 5(1), 73–87*, 2015.

[16] Samson Yoseph Esayas. The role of anonymisation and pseudonymisation under the eu data privacy rules: beyond the 'all or nothing' approach. *in European Journal of Law and Technology, Vol 6, No 2*, 2015.

[17] European Data Protection Board. Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, 2024.

[18] European Data Protection Board. Guidelines on pseudonymisation, 2025.

[19] European General Court. Case T-557/20 SRB, 2023.

[20] European Medicines Agency. European medicines agency policy on publication of clinical data for medicinal products for human use. Policy EMA/144064/2019, European Medicines Agency, 3 2019.

[21] Michèle Finck and Frank Pallas. They who must not be identified—distinguishing personal from non-personal data under the gdpr. *International Data Privacy Law*, 10(1):11–36, 03 2020.

[22] Andrea Gadotti, Luc Rocher, Florimond Houssiau, Ana-Maria Creţu, and Yves-Alexandre De Montjoye. Anonymization: The imperfect science of using data while preserving privacy. *Science Advances 10(29), eadn7053*, 2024.

[23] Garante. Decision PDP [9913795] Registro dei provvedimentin°226 del 1° giugno 2023 1 The THIN database, 2023.

[24] Mostafa Langarizadeh, Azam Orooji, and Abbas Sheikhtaheri. Effectiveness of anonymization methods in preserving patients' privacy: a systematic literature review. *Health Informatics Meets eHealth*, pages 80–87, 2018.

[25] Szivia LestyĂAn, William Letrone, Ludovica Robustelli, and Gergely BiczĂłk. Anonymity-washing. *arXiv preprint arXiv:2505.18627*, 2025.

[26] Alexandre Lodie and Cédric Lauradoux. Is it personal data? solving the gordian knot of anonymisation. In *Privacy Symposium, Venice*, 2024.

[27] Csányi Gergely Márk, Daniel Nagy, Renátó Vági, Pal Vadász, and Tamás Orosz. Challenges and open problems of legal document anonymization. *Symmetry, 13(8):1490*, 2021.

[28] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *IEEE SP*, 2008.

[29] Paul Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review 57, 1701*, 2009.

[30] OWASP, Cheat Sheets Series Team. Password storage cheat sheet, 2025. Accessed August 10, 2025. https://cheatsheetseries.owasp.org/cheatsheets/Password_Storage_Cheat_

Sheet.html.

[31] Nadezhda Purtova. From knowing by name to targeting: the meaning of identification under the gdpr. *International Data Privacy Law, 12*, 2022.

[32] Ira S. Rubinstein and Woodrow Hartzog. Anonymization and risk. *Wash. L. Rev. 91, 703*.

[33] Valentin Rupp and Max von Grafenstein. Clarifying "personal data" and the role of anonymisation in data protection law: Including and excluding data from the scope of the gdpr (more clearly) through refining the concept of data protection. *Computer Law Security Review, Volume 52*, 2024.

[34] Rolf Schwartmann, Andreas Jaspers, Niels Lepperhoff, Steffen Weiß, and Michael Meier. Practice guide to anonymising personal data: Requirements, application classes and procedure model. practice guide, Foundation for Data Protection. Original title: Praxisleitfaden zum Anonymisieren personenbezogener Daten.

[35] Gerald Spindler and Philipp Schmechel. Personal data and encryption in the european general data protection regulation. *IPITEC*, 2016.

[36] Sophie Stalla-Bourdillon. Identifiability, as a Data Risk: Is a Uniform Approach to Anonymisation About to Emerge in the EU? pages 1–19.

[37] Sophie Stalla-Bourdillon and Alison Knight. Anonymous Data v. Personal Data - False Debate: An EU Perspective on Anonymization, Pseudonymization and Personal Data. 34:284.

[38] Latanya Sweeney. k-anonymity: A model for protecting privacy. *IJUFKS 10(5)*, page 557–570, 2002.

[39] Kalliopi Terzidou. Automated anonymization of court decisions: Facilitating the publication of court decisions through algorithmic systems. In *n Nineteenth International Conference for Artificial Intelligence and Law (ICAIL 2023)*, 2023.

[40] Emily Weitzenboeck, Pierre Lison, Malgorzata Agnieszka Cyndecka, and Malcolm Langford. The GDPR and Unstructured Data: Is Anonymization Possible?

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251590*

# Label-Aware Contextual Retrieval for Few-Shot Legal Argument Mining

Huihui Xu [a,b,1], Kevin Ashley [a,b,c]
[a] *Intelligent Systems Program, University of Pittsburgh*
[b] *Learning Research and Development Center, University of Pittsburgh*
[c] *School of Law, University of Pittsburgh*

**Abstract.** Two major challenges for legal argument mining are limited annotated data and the need to model context-dependent reasoning. This paper presents a framework that enhances few-shot in-context learning (ICL) for legal argument mining by combining context-augmented similarity with hard label diversity filtering. Our method selects example sentences not only based on their semantic closeness to the target but also by incorporating their surrounding context and ensuring diverse label representation. Our legal argument mining labels are Issue, Reason, Conclusion, and Non-IRC classes. We evaluate the interaction between different context configurations (pre-, post-, and bidirectional) and label diversity under various few-shot settings on a legal case summary annotation task. Results show that our combined approach outperforms traditional retrieval strategies across multiple large language models with improved macro F1 scores. This work demonstrates a way of improving legal argument mining in low-resource settings by optimizing the selection of few-shot examples.

**Keywords.** argument mining, in-context learning, information retrieval, large language models

## 1. Introduction

Legal case documents often contain specific argumentative components such as **Issue**s, **Reason**s, and **Conclusion**s, commonly referred to as **IRC**s [1,2]. Legal argument mining aims to automatically identify and label sentence roles such as these in case texts to support downstream tasks like legal summarization [1,3] and legal question answering [4]. Despite growing interest in this task, annotated datasets for IRC labeling remain limited because of the high cost and expertise for annotation.

Recent advances in large language models (LLMs) have made in-context learning (ICL) a promising alternative to full-scale fine-tuning. ICL involves providing a manually crafted prompt, which may include a task description and a small set of demonstration examples [5]. A few example inputs are given, and ICL combines them with the test example as input. LLMs perform inference conditioned on the demonstrations. ICL reduces the need for high-quality, human-expert annotated data, where annotations are costly and difficult to obtain [6]. Recent studies have explored ICL-based approaches to

[1]Corresponding Author: Huihui Xu; Email: huihui.xu@pitt.edu

argument mining. For example, [7] combine k-NN-based example retrieval with ensembling to perform argument type classification. In the legal domain, [8] demonstrate that GPT-3.5 and GPT-4's performance on classifying argument components highly depends on prompt structure and example selection. To alleviate the issue, several methods have been proposed to select examples for few-shot learning in order to optimize the model performance [9].

While prior work in argument mining has leveraged ICL, it often treats demonstrations as isolated units, overlooking two key challenges in legal argument mining: (1) the need to account for sentence-level context, which is essential for interpreting the argumentative function of legal text, and (2) the risk of label imbalance, where majority class examples dominate few-shot prompts. This is problematic in legal texts, since legal argument components like Issue, Reason and Conclusion are often interdependent because the value of one example comes from contrasting it with others. In IRCs, if one type dominates, those contrasts disappear, so the model will over rely on the repeated type and lead to a wrong label.

Our approach aims to address the two challenges mentioned above by investigating both surrounding context and label diversity to improve ICL performance for legal argument mining. Our contributions include:

- A novel context-augmented similarity mechanism that incorporates pre-, post-, and bidirectional context during retrieval to select more informative examples of legal argument types.
- A label-aware diversity filtering strategy that ensures balanced coverage of argumentative types in few-shot prompts.
- Empirical evidence that combining context and label diversity leads to performance improvements over standard semantic or diversity-based retrieval baselines.

## 2. Related Work

### *2.1. Legal Argument Mining*

Legal argument mining aims to automatically identify and classify argumentative components. Early work in this area focused on supervised learning with extensive expert annotated datasets. [1,2] developed Issue-Reason-Conclusion (IRC) scheme for argument representation and experimented different deep learning models for supervised label classification. [10] developed a dataset of Indian Supreme Court judgments labeled with rhetorical roles, and proposed early sequence models for classification. [11] introduces SCALE, a semi-supervised case annotation for legal explanations using embeddings. [12] proposed a new annotation scheme and curated an expert-annotated dataset from European Court of Human Rights (ECHR) decisions. They fine-tuned BERT-based models to improve the classification performance.

Recent works has investigated using ICL with Large Language Models (LLMs) to reduce annotation costs across legal tasks, such as legal explanatory values annotation [13] and legal factor extraction [14]. [8] used GPT-3.5 and GPT-4 for classifying argument components. [7] further explored the few-shot ICL examples selection with LLMs.

Most of the prior work treats few-shot examples as independent inputs without considering the importance of sentence-level context and the structural interdependence between argumentative components. To address these limitations, we introduce a context-augmented, label-aware example selection strategy tailored for few-shot legal argument mining. Our method enhances example selection by leveraging sentence-level context and enforcing label diversity constraints. It produces in-context prompts that have better coverage of the task's full argumentative space.

### *2.2. Example Selection for In-Context Learning*

LLMs have shown great zero-shot performance across various NLP tasks [15,16,17,18]. Research shows that adopting ICL with a few examples can greatly improve model performance on new tasks [17,5,19,20,9,21,22]. This capability stems from LLMs'extensive pre-training on diverse datasets and opens a lot of opportunities for leveraging LLMs in more dynamic applications.

As noted, despite the powerful ICL capability, research has shown that the performance of ICL is sensitive to the selection of the demonstration samples. [20] investigated the sensitivity of GPT-3 to in-context samples. They proposed a strategy to select good examples for in-context learning and demonstrated the effectiveness of retrieving in-context examples in terms of the semantic similarity to the test samples. [9] explored example selection by quantifying the ICL effect of a given ICL example. They calculated influence scores on an additional hold-out set. This method requires intensive computation when the pool of candidate examples is large. [21] proposed a two-stage frame work for self-selection demonstration examples by applying random algorithms. They first retrieve candidates that are informative for ICL and determine their optimal ordering and compression. [22] proposed a framework to quantify the uncertainty in in-context learning by decomposing it into aleatoric and epistemic components based on the generated outputs.

Prior methods were often tailored to tasks like topic classification or semantic classification in general-purpose domains, but do not adequately adapt to the nuances of ICL within the legal domain. Legal argument mining presents distinct challenges: legal argument components are often embedded in complex rhetorical structures and depend heavily on contextual interpretation. Additionally, the uneven distribution of these labels in legal texts introduces majority-class bias when selecting few-shot examples.

To address these domain specific challenges, we propose this context-augmented, label-aware example selection strategy. Our approach differs from purely semantic or influence-based retrieval by incorporating the surrounding sentence context (pre-, post-, or bidirectional context) into the similarity computation. Furthermore, we introduce a label-diversity filtering step to ensure broader coverage of underrepresented argumentative types to mitigate the bias that often arises in few-shot prompts.

## 3. Problem Formulation

Legal argument identification often depends on the broader context surrounding a given sentence. To capture this, we construct a context-augmented representation for each target sentence $s_i$, which includes a configurable number of neighboring sentences before or after the target sentence. We represent the structure formally as following:

$$\text{Context}(s_i) = [\text{PRE}_k, \text{PRE}_{k-1}, \ldots, \text{PRE}_1, s_i, \text{POST}_1, \ldots, \text{POST}_k] \quad (1)$$

Here, $\text{PRE}_j$ and $\text{POST}_j$ correspond to the $j$-th preceding and following sentences, respectively. The context window size $k$ and directionality (pre-only, post-only, or both) are adjustable parameters in our design. This formulation allows us to systematically assess the context impact for few-shot ICL learning.

While incorporating contextual information improves the relevance and informativeness of candidate examples, semantic similarity alone is insufficient to guarantee balanced label representation. To mitigate this issue, we introduce a hard label-aware filtering mechanism that enforces class diversity within the selected demonstration set.

Starting with a ranked list of candidates $C = c_1, c_2, \ldots, c_{3k}$ obtained via semantic similarity or Maximal Marginal Relevance (MMR), we then iteratively select the top-$k$ examples by including only those examples that introduce a previously unseen label if possible. If fewer than $k$ unique labels are presented in the candidate pool, we fill the remaining slots with the most relevant examples available. This approach ensures that the final few-shot prompt contains a broad range of legal argument types while preserving semantic alignment with the target instance.

## 4. Methodology

We implement our selection strategy through a multi-step pipeline that combines context augmentation with label-aware filtering. The process begins by augmenting both query and candidate examples with surrounding context, followed by retrieval using semantic similarity or MMR. We then apply a hard label filtering step to promote class diversity in the final in-context prompt.

### *4.1. Representation Learning with Contextual Information*

Unlike standard sentence classification tasks, legal argument identification often depends on cues from the surrounding textual context. For each target sentence $s_i$ in a document $D = \{s_1, s_2, \ldots, s_n\}$, we extract up to $w$ neighboring sentences based on a configurable context window. To ensure robustness at document boundaries, the algorithm verifies index validity before accessing context sentences. We define four context configurations to systematically evaluate the effect of directionality:

- **Pre-context:** Includes the $k$-th preceding sentence, marked as [`PRE_CONTEXT_k`].
- **Post-context:** Includes the $k$-th following sentence, marked as [`POST_CONTEXT_k`].
- **Bidirectional:** Combines both preceding and following context with corresponding markers.
- **No-context (baseline):** Uses only the target sentence $s_i$ without additional context.

The fully formatted context-augmented string is processed as a single input to the SBERT model (`all-MiniLM-L6-v2`). This enables the model to capture attention across the entire sequence, while different context configurations supply varying amounts of surrounding information. Embeddings are computed as:

$$e_i = \text{SBERT}(\text{format}(\text{context}(s_i))) \in R^{384} \quad (2)$$

where format$(\cdot)$ denotes the application of context markers and concatenation rules according to the selected configuration. Window sizes are held constant within each configuration to ensure consistent context length.

Context augmentation changes the similarity computation by extending the representation space from isolated sentences to include surrounding legal discourse. Each context configuration defines a distinct semantic perspective: pre-context emphasizes preceding argumentative flow, post-context highlights subsequent elaboration, and bidirectional context captures balanced surrounding cues.

To compute the similarity between a query $q$ and a candidate $x_i$, we apply cosine similarity over their context-augmented embeddings:

$$\text{sim}_{\text{context}}(q, x_i) = \frac{E(\text{context}(q)) \cdot E(\text{context}(x_i))}{\|E(\text{context}(q))\| \cdot \|E(\text{context}(x_i))\|} \tag{3}$$

where $E(\cdot)$ denotes the embedding function applied to the formatted, context-augmented input. This context-aware similarity captures not only semantic alignment between the query and candidate, but also the surrounding legal reasoning patterns encoded through fixed window configurations—enabling controlled analysis of contextual effects on retrieval.

### *4.2. In-Context Learning Example Retrieval Methods*

We benchmark our proposed method against several baseline retrieval strategies that reflect common practices in few-shot example selection.

The *semantic similarity* baseline performs nearest-neighbor retrieval to select training examples most similar to the query sentence. Given a query $q$ and a labeled training set $T = \{(x_i, y_i)\}_{i=1}^{N}$, we compute cosine similarity scores between $q$ and each candidate $x_i$. To enable efficient retrieval and ranking, we employ FAISS (Facebook AI Similarity Search) [23]. This method yields a ranked candidate pool $C_{\text{semantic}} = \{c_1, c_2, \ldots, c_{3k}\}$, ordered by decreasing similarity to the query. This approach prioritizes semantic relevance but may result in redundant examples covering similar topics or over-representing certain label classes.

The *Maximal Marginal Relevance* (MMR) [24] method addresses the redundancy limitations of pure similarity-based retrieval by jointly optimizing for both relevance to the query and diversity among selected examples. At each iteration, the next example $x_i$ is chosen to maximize the following objective:

$$\text{MMR}(x_i) = \lambda \cdot \text{sim}(q, x_i) - (1 - \lambda) \cdot \max_{x_j \in S} \text{sim}(x_i, x_j) \tag{4}$$

where $S$ denotes the set of already selected examples, and $\lambda$ is a tunable parameter controlling the trade-off between relevance and diversity (set to $\lambda = 0.5$ in our experiments). The algorithm begins by selecting the most relevant example and then iteratively adds examples that are both similar to the query and dissimilar to those already chosen. The resulting candidate pool $C_{\text{mmr}}$ exhibits greater topical and lexical diversity than that produced by semantic similarity alone.

Base retrieval methods prioritize semantic relevance but often overlook class balance. It leads to few-shot prompts biased toward majority labels. To address this, we apply a hard label-aware filtering strategy that promotes label diversity without sacrificing relevance. Given a ranked candidate pool $C = \{c_1, c_2, \ldots, c_{3k}\}$, we iteratively select $k$ examples by greedily adding those with unseen labels to ensure a more balanced and representative prompt. The selection procedure is described in Algorithm 1. An example is added only if its label has not been selected. If fewer than $k$ unique labels are available, the remaining slots are filled with the next most relevant examples regardless of label.

**Algorithm 1** Context-Augmented Hard Label-Aware Filtering

**Require:** Query $q$, training set $D = \{(x_i, y_i)\}_{i=1}^{N}$, number of shots $k$, embedding model $E$

**Ensure:** Selected examples $S = \{(x_j, y_j)\}_{j=1}^{k}$

**Step 1: Context-Augmented Similarity Retrieval**

Compute $e_q \leftarrow E(q)$
**for** $i = 1$ to $N$ **do**
    $e_i \leftarrow E(\texttt{combined_input}(x_i))$ ▷ Includes optional context
**end for**
Rank candidates using semantic similarity or MMR
$C \leftarrow \{(x_j, y_j)\}_{j=1}^{3k}$

**Step 2: Hard Label-Aware Filtering**

Initialize $S \leftarrow [\,]$, label set $L \leftarrow \emptyset$
**for** each $(x_j, y_j)$ in $C$ **do**
    **if** $y_j \notin L$ **then**
        Add $(x_j, y_j)$ to $S$; add $y_j$ to $L$
    **end if**
    **if** $|S| = k$ **then**
        **break**
    **end if**
**end for**
**if** $|S| < k$ **then**
    **for** each $(x_j, y_j)$ in $C$ **do**
        **if** $(x_j, y_j) \notin S$ **then**
            Add $(x_j, y_j)$ to $S$
        **end if**
        **if** $|S| = k$ **then**
            **break**
        **end if**
    **end for**
**end if**
**return** $S$

## 5. Experimentation

### 5.1. Dataset and Models

We experiment with a dataset of human-written legal case summaries annotated for argumentative elements. Every sentence in a summary is labeled with one of four possible labels: Issue (the legal question under dispute), Reason (the supporting rationale or legal reasoning), Conclusion (the court's conclusion to the issue), or Non IRC (non-argumentative content). The dataset consists of 171 legal cases in total, with 149 cases (1646 sentences) used for training and 22 cases (202 sentences) for testing.

We test our proposed method with three open-source models: LLaMA3.1-70b, LLaMA3.3-70b [25], Cogito-v1-Preview-70b [26]. All models are deployed using Ollama, an open-source framework for LLM management and deployment. All models are quantized to 4-bit precision (Q4) to reduce the memory and computational overhead. It can significantly minimizes memory usage while maintaining relatively high model performance. The deployment leverages NVIDIA L40s GPUs with 48GB memory, driver version 550.54.15, and CUDA version 12.4.

### 5.2. Experimental Configurations

We set the context window sizes to $w_{\text{pre}} = w_{\text{post}} = 2$ for the pre-context and post-context configurations, and $w_{\text{bidirectional}} = 1$ for the bidirectional context (i.e., one preceding and one following sentence). The choice of window size depends on the structure of legal summary and computational efficiency. Context markers follow the notation `[PRE_CONTEXT_k]` and `[POST_CONTEXT_k]` to maintain consistent formatting across all experimental designs.

For base retrieval methods, we generate candidate pools of size $3k$, where $k$ represents the target number of ICL examples. We experiment with $k \in \{2,4\}$ to match the number of possible argument labels in our tasks. We employ SBERT (`all-MiniLM-L6-v2`) for text embedding, which producing 384-dimensional output dense embedding vectors.

We set temperature $T = 0$ for all LLM inference to ensure deterministic outputs for reproducibility. We implement structured output parsing using the `instructor` library to extract legal argument labels in a consistent format. The codebase is available at Github.

## 6. Results and Discussion

### 6.1. How does context-augmented example selection compare to traditional text-only similarity approaches in classifying argumentative roles within legal case summaries?

**Result**

Table 1 presents the performance of three language models across four context configurations (no context, pre-only, post-only, bidirectional) and two retrieval strategies (semantic similarity and MMR). Overall, incorporating context consistently improved F1 scores across models and retrievers compared to the no-context baseline. Bidirectional context yielded the highest gains in IRC categories, with Llama3.1-70b achieving a peak

overall F1 of 0.74 (4-shot) using semantic retrieval. Context effects were also more pronounced in higher k-shot settings, where most 4-shot configurations outperform their 2-shot counterparts. Among retrievers, the MMR retriever slightly outperformed semantic for the Issue and Conclusion categories, although the margin was narrow.

**Discussion**

We observe that incorporating contextual sentences into MMR-based example selection leads to measurable gains in overall F1 scores. These findings confirm that incorporating contextual information can improve few-shot legal argument mining performance. Context augmentation is especially helpful for context-sensitive labels like Issue (F1 score jumps from 0.67(no context) to 0.80 (bidirectional) for semantic retriever ). Overall, the results highlight the importance of rhetorical context and when selecting effective in-context examples for legal tasks.

**Table 1.** Comparison of context-augmented and non-context example selection across different models, retrievers, and context window configurations. Each cell reports 4-shot performance, and 2-shot performance are in parentheses. **Bolded** values represent the best performance for 4-shot prompting across all settings, while underlined values indicate the best performance for 2-shot prompting. F1 refers to the macro-F1 score

| Model | Context | Retriever | Issue | Reason | Conclusion | Overall | | |
|---|---|---|---|---|---|---|---|---|
| | | | F1 | F1 | F1 | P | R | F1 |
| Llama3.1-70b | No Context | Semantic | .67 (.62) | .71 (.70) | .78 (.76) | .72 (.70) | .72 (.71) | .71 (.70) |
| | | MMR | .66(.63) | .71 (.70) | .80 (.81) | .71 (.71) | .71 (.72) | .70 (.71) |
| | Pre-only | Semantic | .71 (.68) | **.72** (.73) | .78 (.78) | .76 (.74) | .73 (.72) | .74 (.72) |
| | | MMR | .78 (.70) | .68 (.72) | .77 (.75) | .75 (.74) | .73 (.72) | .74 (.72) |
| | Post-only | Semantic | .75 (.69) | .65(.63) | .77(.72) | .75 (.70) | .72 (.67) | .72 (.68) |
| | | MMR | .77 (.71) | .68 (.69) | .73 (.75) | .74 (.73) | .72 (.70) | .72 (.71) |
| | Bidirectional | Semantic | .80 (.76) | .67 (.67) | **.81** (.78) | .76 (.75) | **.74** (.72) | **.74** (.73) |
| | | MMR | **.81** (.76) | .65 (.66) | .79 (.74) | **.77** (.75) | .74 (.71) | .74 (.71) |
| Llama3.3-70b | No Context | Semantic | .61 (.62) | .67 (.66) | .80 (.76) | .70 (.69) | .71 (.70) | .70 (.69) |
| | | MMR | .64 (.57) | .68 (.65) | .78 (.77) | .70 (.67) | .71 (.68) | .70 (.67) |
| | Pre-only | Semantic | .69 (.68) | **.73** (.72) | .80 (.74) | **.73** (.72) | **.74** (.72) | .73 (.71) |
| | | MMR | .71 (.64) | .72 (.70) | .80 (.74) | .73 (.71) | .74 (.70) | .73 (.70) |
| | Post-only | Semantic | .65 (.62) | .65 (.61) | .77 (.72) | .70 (.66) | .71 (.67) | .70 (.66) |
| | | MMR | **.73** (.68) | .67 (.63) | .78 (.77) | .72 (.70) | .73 (.70) | .72 (.69) |
| | Bidirectional | Semantic | .72 (.77) | .65 (.62) | .76 (.75) | .71 (.72) | .71 (.72) | .71 (.72) |
| | | MMR | .73 (.75) | .66 (.62) | **.84** (.79) | **.73** (.72) | .73 (.72) | **.73** (.72) |
| Cogito-v1-70b | No Context | Semantic | .63 (.60) | .68 (.62) | .77 (.78) | .70 (.69) | .69 (.69) | .69 (.68) |
| | | MMR | .62 (.51) | .70 (.64) | .82 (.79) | .72 (.68) | .72 (.67) | .72 (.66) |
| | Pre-only | Semantic | .70 (.63) | .69 (.62) | .83 (.71) | **.74** (.69) | .73 (.68) | .73 (.68) |
| | | MMR | .69 (.58) | **.72** (.69) | .80 (.78) | .73 (.70) | **.74** (.69) | .73 (.69) |
| | Post-only | Semantic | .66 (.62) | .67 (.59) | .81 (.77) | .72 (.68) | .72 (.67) | .72 (.67) |
| | | MMR | .61 (.60) | .66 (.62) | .78 (.78) | .69 (.69) | .70 (.68) | .69 (.67) |
| | Bidirectional | Semantic | **.77** (.64) | .67 (.60) | .79 (.78) | **.74** (.69) | .74 (.68) | .74 (.68) |
| | | MMR | .75 (.60) | .66 (.61) | **.83** (.79) | **.74** (.68) | **.74** (.68) | **.74** (.68) |

*6.2. Does contextual augmentation improve the effectiveness of few-shot prompts when combined with hard label-aware diversity filtering in legal argument mining?*

**Result**

Table 2 shows that combining contextual augmentation with hard label-aware diversity filtering consistently improves performance across all models and label types. Each cell reports 4-shot performance with 2-shot results shown in parentheses.

While contextual augmentation consistently improves performance, optimal strategies vary across different models. Llama3.1-70b achieves its highest overall F1 (0.76) with semantic and pre-only context, and Llama3.3-70b performs best with Semantic with bidirectional, achieving the highest overall F1 (0.73). Cogito-v1-70b reaches 0.74 overall F1 with both semantic with bidirectional and MMR with pre-only.

In some cases, 2-shot prompting outperforms 4-shot, such as Llama3.1-70b on Issue (F1 = 0.84 with MMR and bidirectional context). It suggests that fewer but well-selected examples maybe more effective.

**Discussion**

The results show that context generally boosts performance across models, but the best configuration varies. For instance, semantic with pre-only context performs best for Llama3.1-70b, while semantic with bidirectional context is optimal for Llama3.3-70b. Besides, semantic with bidirectional benefits more under the label-filtering condition. It suggests that contextual signals are more effectively leveraged when semantically similar examples are selected by label-aware retrieval. MMR-based retrieval does not benefit as much because it has already optimized for diversity during example selection. It reduces the marginal gains from the filtering strategy.

*6.3. Does hard label-aware diversity filtering amplify the benefits of contextual retrieval compared to unfiltered example selection strategies in legal argument classification?*

**Result**

We observe that hard label-aware diversity filtering amplifies the benefits of contextual retrieval, especially for the semantic-based retriever. For example, with 4-shot prompting, Llama3.1-70b with semantic retrieval and pre-context improves from an overall F1 of 0.74 (unfiltered) to 0.76 (filtered). For individual classification, filtering amplifies performance gains for Issue and Reason labels more than for Conclusion. For example, Cogito-v1-70b with semantic retrieval and label-aware filtering achieves the highest Issue F1 score of 0.79, which outperform its unfiltered counterpart (F1 = 0.77). This trend holds across both 4-shot and 2-shot prompting.

**Discussion**

These results suggest that label-aware filtering and contextual retrieval are synergistic, but the extent of benefit depends on the base retriever. Semantic retrieval gains more from label-aware filtering because it corrects for label imbalances of underrepresented categories. In contrast, MMR already enforces diversity during the sample selection, which limits the marginal utility of additional label filtering. In general, our findings confirmed that filtering is most effective when paired with a weaker retriever.

## 7. Conclusion

This paper introduces a novel context-augmented and label-aware example selection strategy to enhance few-shot in-context learning for legal argument mining. By integrat-

**Table 2.** We evaluate two base retrievers (*MMR label-ware filtering* and *Semantic label-ware filtering*) with varying contextual configurations. Each cell reports result under 4-shot prompting, with corresponding 2-shot results shown in parentheses. **Bolded** values represent the best performance for 4-shot prompting across all settings, while underlined values indicate the best performance for 2-shot prompting. F1 refers to the macro-F1 score.

| Model | Retriever | Context | Issue | Reason | Conclusion | Overall | | |
|---|---|---|---|---|---|---|---|---|
| | | | F1 | F1 | F1 | P | R | F1 |
| Llama3.1-70b | Semantic + Filtering | No context | .68 (.64) | **.73** (.71) | .76 (.76) | .71 (.71) | .72 (.72) | .71 (.71) |
| | | Pre-only | .80 (.71) | .72 (.70) | .78 (.76) | **.78** (.75) | **.76** (.72) | **.76** (.72) |
| | | Post-only | .76 (.69) | .66 (.64) | .77 (.74) | .76 (.72) | .72 (.69) | .73 (.69) |
| | | Bidirectional | **.81** (.77) | .67 (.66) | .77 (.74) | **.78** (.73) | .74 (.71) | .74 (.71) |
| | MMR + Filtering | No context | .66 (.60) | .72 (.70) | **.79** (.81) | .72 (.70) | .72 (.71) | .72 (.71) |
| | | Pre-only | .75 (.70) | .67 (.71) | .74 (.74) | .74 (.74) | .72 (.71) | .72 (.71) |
| | | Post-only | .74 (.73) | .67 (.69) | .73 (.77) | .72 (.74) | .71 (.72) | .70 (.72) |
| | | Bidirectional | .77 (.74) | .69 (.65) | .74 (.71) | .75 (.73) | .72 (.68) | .72 (.69) |
| Llama3.3-70b | Semantic + Filtering | No context | .59 (.55) | .66 (.64) | .75 (.75) | .67 (.66) | .68 (.66) | .67 (.65) |
| | | Pre-only | .69 (.65) | **.77** (.70) | .80 (.76) | .74 (.72) | .75 (.70) | .74 (.71) |
| | | Post-only | .67 (.66) | .65 (.64) | .77 (.75) | .70 (.69) | .70 (.70) | .70 (.69) |
| | | Bidirectional | **.79** (.78) | .65 (.63) | .80 (.77) | **.73** (.73) | **.75** (.73) | **.73** (.73) |
| | MMR + Filtering | non-context | .63 (.53) | .67 (.62) | .77 (.76) | .70 (.65) | .71 (.66) | .70 (.65) |
| | | Pre-only | .66 (.64) | .72 (.72) | .81 (.77) | .71 (.72) | .72 (.71) | .71 (.71) |
| | | Post-only | .73 (.67) | .70 (.66) | .77 (.78) | .72 (.70) | .73 (.71) | .73 (.70) |
| | | Bidirectional | .69 (.73) | .67 (.61) | **.82** (.77) | .72 (.71) | .72 (.71) | .71 (.71) |
| Cogito-v1-70b | Semantic + Filtering | No context | .61 (.54) | .69 (.62) | .79 (.77) | .69 (0.66) | .70 (0.65) | .69 (0.65) |
| | | Pre-only | .63 (.58) | **.77** (.65) | **.82** (.71) | **.74** (0.68) | .73 (0.65) | .73 (0.66) |
| | | Post-only | .63 (.55) | .67 (.62) | .79 (.78) | .71 (0.67) | .70 (0.66) | .70 (0.66) |
| | | Bidirectional | **.79** (.71) | .63 (.60) | .81 (.79) | **.74** (0.71) | **.75** (0.70) | **.74** (0.70) |
| | MMR + Filtering | No context | .56 (.50) | .66 (.63) | .77 (.77) | .69 (0.66) | .68 (0.66) | .68 (0.65) |
| | | Pre-only | .69 (.56) | .71 (.67) | **.82** (.77) | .73 (0.68) | .74 (0.66) | **.74** (0.67) |
| | | Post-only | .62 (.63) | .65 (.62) | .80 (.78) | .69 (0.70) | .69 (0.68) | .69 (0.68) |
| | | Bidirectional | .68 (.61) | .62 (.61) | .78 (.79) | .70 (0.68) | .69 (0.67) | .69 (0.67) |

ing sentence-level context and enforcing label diversity, our method addresses limitations of prior ICL approaches that often overlook the structural dependencies and class imbalance. Experimental results across multiple models show that combining contextual retrieval with label-aware filtering improves performance in identifying argumentative roles.

## 8. Limitations and Future Work

Our work has the following limitations. First, the evaluation relies on a finite and relatively small set of annotated legal summaries. It might limit generalizability to broader domains. Second, we simulate few-shot settings with a fixed prompt design, and performance may be sensitive to prompt design. Additionally, the current study assumes uniform quality across retrieved examples and relies on predefined label annotations without accounting for edge cases. Finally, while we conducted several ablations, we did not explore certain factors (e.g., context window size, candidate pool size, or quantization levels) due to space and computational constraints, which we leave for future work.

In future work, we plan to expand our work to more diverse label datasets for evaluating the robustness of the label-aware contextual ICL system.

## Acknowledgments

This work was supported by the National Science Foundation (Grant No. 2040490, FAI: Using AI to Increase Fairness by Improving Access to Justice) and by a Pitt Momentum Funds Scaling Grant. This research was supported in part by the University of Pittsburgh Center for Research Computing and Data, RRID:SCR_022735, through the resources provided. Specifically, this work used the H2P cluster, which is supported by NSF award number OAC-2117681. We thank the CRCD team, particularly Yassin Khalifa, PhD, Chengnian Xiao, PhD, and Fangping Mu, PhD, for their assistance with computing resources.

## References

[1] Xu H, Šavelka J, Ashley KD. Using argument mining for legal text summarization. In: Legal knowledge and information systems. IOS Press; 2020. p. 184-93.
[2] Xu H, Ashley K. Multi-Granularity Argument Mining in Legal Texts. In: Legal Knowledge and Information Systems. vol. 362 of Frontiers in Artificial Intelligence and Applications. IOS Press; 2022. p. 261-6.
[3] Xu H, Ashley K. Argumentative Segmentation Enhancement for Legal Summarization. 2023;3441:141-50. Available from: https://ceur-ws.org/Vol-3441/paper15.pdf.
[4] Xu H, Ashley K. A Question-Answering Approach to Evaluating Legal Summaries. In: Legal Knowledge and Information Systems. IOS Press; 2023. p. 293-8.
[5] Brown TB, Mann B, Ryder N, Subbiah Mea. Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20. Red Hook, NY, USA: Curran Associates Inc.; 2020. .
[6] Wu X, Xiao L, Sun Y, Zhang J, Ma T, He L. A survey of human-in-the-loop for machine learning. Future Generation Computer Systems. 2022;135:364-81.
[7] Cabessa J, Hernault H, Mushtaq U. In-context learning and fine-tuning GPT for argument mining. arXiv preprint arXiv:240606699. 2024.
[8] Al Zubaer A, Granitzer M, Mitrović J. Performance analysis of large language models in the domain of legal argument mining. Frontiers in artificial intelligence. 2023;6:1278796.
[9] Nguyen T, Wong E. In-context example selection with influences. arXiv preprint arXiv:230211042. 2023.
[10] Bhattacharya P, Paul S, Ghosh K, Ghosh S, Wyner A. Identification of rhetorical roles of sentences in indian legal judgments. In: Legal knowledge and information systems. IOS Press; 2019. p. 3-12.
[11] Branting LK, Pfeifer C, Brown B, Ferro L, Aberdeen J, Weiss B, et al. Scalable and explainable legal prediction. Artificial Intelligence and Law. 2021;29(2):213-38.
[12] Habernal I, Faber D, Recchia N, Bretthauer S, Gurevych I, Spiecker genannt Döhmann I, et al. Mining legal arguments in court decisions. Artificial Intelligence and Law. 2024;32(3):1-38.
[13] Savelka J, Ashley K, Gray M, Westermann H, Xu H. Can GPT-4 Support Analysis of Textual Data in Tasks Requiring Highly Specialized Domain Expertise? In: Proceedings of the 6th Workshop on Automated Semantic Analysis of Information in Legal Text Co-Located with the 19th International Conference on Artificial Intelligence and Law (ICAIL 2023); 2023. Available from: https://ceur-ws.org/Vol-3441/paper1.pdf.
[14] Gray MA, Savelka J, Oliver WM, Ashley KD. Empirical legal analysis simplified: reducing complexity through automatic identification and evaluation of legally relevant factors. Philosophical Transactions of the Royal Society A. 2024;382(2270):20230155.
[15] Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22. Red Hook, NY, USA: Curran Associates Inc.; 2024. .
[16] Zhang K, Gutiérrez BJ, Su Y. Aligning Instruction Tasks Unlocks Large Language Models as Zero-Shot Relation Extractors. In: Findings of the Association for Computational Linguistics: ACL 2023; 2023. p. 794-812.

[17] Labrak Y, Rouvier M, Dufour R. A Zero-shot and Few-shot Study of Instruction-Finetuned Large Language Models Applied to Clinical and Biomedical Tasks. In: Calzolari N, Kan MY, Hoste V, Lenci A, Sakti S, Xue N, editors. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Torino, Italia: ELRA and ICCL; 2024. p. 2049-66. Available from: https://aclanthology.org/2024.lrec-main.185.

[18] Fazili B, Agrawal A, Jyothi P. Boosting Zero-Shot Crosslingual Performance using LLM-Based Augmentations with Effective Data Selection. In: Ku LW, Martins A, Srikumar V, editors. Findings of the Association for Computational Linguistics: ACL 2024. Bangkok, Thailand: Association for Computational Linguistics; 2024. p. 13406-22. Available from: https://aclanthology.org/2024.findings-acl.795.

[19] Wei J, Wang X, et al S. Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of the 36th International Conference on Neural Information Processing Systems; 2022. p. 24824-37.

[20] Liu J, Shen D, Zhang Y, Dolan Bea. What Makes Good In-Context Examples for GPT-3? In: Agirre E, Apidianaki M, Vulić I, editors. Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures. Dublin, Ireland and Online: Association for Computational Linguistics; 2022. p. 100-14. Available from: https://aclanthology.org/2022.deelio-1.10.

[21] Wu Z, Wang Y, Ye J, Kong L. Self-Adaptive In-Context Learning: An Information Compression Perspective for In-Context Example Selection and Ordering. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2023. p. 1423-36.

[22] Ling C, Zhao X, et al Z. Uncertainty Quantification for In-Context Learning of Large Language Models. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers); 2024. p. 3357-70.

[23] Douze M, Guzhva A, Deng C, Johnson J, Szilvasy G, Mazaré PE, et al. The Faiss library. 2024.

[24] Carbonell J, Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval; 1998. p. 335-6.

[25] AI@Meta. Llama 3 Model Card. 2024. Available from: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[26] Cogito D. Cogito v1 Preview: Iterated Distillation and Amplification for Hybrid Reasoning; 2025. Accessed July 2025.
urlhttps://www.deepcogito.com/research/cogito-v1-preview.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251591*

# A Multi-Agent AI Framework for Legal Aid: Real-World Interactive Dialogue in Case Consultations

Ying-chu Yu [a] Sieh-chuen Huang [a] Hsuan-lei Shao [b,1]
[a] *College of Law, National Taiwan University, Taipei, Taiwan*
[b] *Graduate Institute of Health and Biotechnology Law, Taipei Medical University*

**Abstract.** This paper presents a multi-agent AI framework for legal aid, designed to support real-world case fact management through interactive dialogue. Our agent framework emulates the iterative questioning, clarification, and synthesis processes of legal professionals, not solely on the fragmented information initially provided by litigants. By engaging in multi-turn interactions, the system incrementally supplements missing details and mitigates risks of misinterpretation, thereby aligning more closely with the dynamics of real legal consultations. A key contribution of this work is the use of real-world legal aid case records as training and evaluation material, ensuring that the framework is grounded in authentic data rather than synthetic simulations. The system is implemented as a collaboration among specialized agents: (1) litigantTwins, which maintains factual integrity and guards against hallucinations; (2) legalAider, which leverages legal knowledge to generate context-sensitive follow-up questions and update case narratives; and (3)Evaluator, which compares AI-generated case records against ground truth facts, assessing factual correctness through qualitative and quantitative measures. Technically, the framework can be built on LLMs, with multi-agent system and prompt design enabling robust coordination. This architecture enhances the quality of fact construction and reasoning, while also offering a scalable solution for online public legal aid services. Beyond its technical contributions, this research highlights its public value, accessibility, and commitment to fairness in the distribution of legal resources. By integrating multi-agent AI with real-world case data, the framework addresses both the technological and socio-legal dimensions of legal information systems, advancing on legal knowledge management, deployment of conversational agents, and normative reasoning in multi-agent systems.

**Keywords.** Multi-Agent AI, Large Language Models (LLMs), Legal Aid, Legal Case Fact Management, Interactive Dialogue Systems

## 1. Introduction

### *1.1. The LLMs of Legal Informatics*

Artificial intelligence has been increasingly applied in the legal domain, ranging from document classification and information retrieval to contract analysis and question-

[1]Corresponding Author: Hsuan-Lei Shao, Ph.D., Associate Professor, ORCID: 0000-0002-7101-5272. Email: hlshao@tmu.edu.tw

answering systems. While these applications demonstrate significant progress, they often share two common limitations. Most systems rely on **single-turn inputs**, producing answers or summaries directly from the initial information provided by litigants. As a result, they fail to capture the iterative and clarifying nature of real legal consultations, where lawyers repeatedly probe, question, and refine narratives to uncover the full scope of facts. In addition, many studies depend on **synthetic or simulated data**, which simplifies the evaluation process but underrepresents the complexity, incompleteness, and ambiguity of real-world case records. These gaps limit the practical utility of current legal AI systems, especially in settings that require close interaction and fact reconstruction.

Access to justice remains a pressing global challenge[17, 19], particularly in contexts where legal aid and legal clinics serve large populations with limited resources. The demand for these services is high, yet the availability of qualified professionals and institutional capacity often falls short. This mismatch highlights an urgent need for innovative solutions that can both support legal practitioners and improve the accessibility of justice for disadvantaged groups. For increasing volumes of information and the desire for more efficient legal processes [1, 2]. These developments enable predictive technologies to forecast litigation outcomes [3, 4]. Furthermore, visualization techniques have become bridging the gap between human understanding and machine processing [5].

legal models to improve the existing legal system[6]. Comprehensive overviews of legal informatics, such as the work by Katz and Dolin [7], provide valuable insights into real-world applications like document review and online dispute resolution.

Another challenge is that LLMs must adapt to diverse legal frameworks, a task that general models are currently ill-equipped to handle[8, 9]. But, loss of legal diversity can be significant. The LLMs often do not notice legal specificities in each state[10, 11]. To address this, there is a growing push for developing customized LLMs that are trained on localized data sets, ensuring that the legal advice and documentation generated are relevant and compliant with local laws[12, 13].

In response, this study introduces a novel **AI-agent framework** that emulates the questioning, clarification, and synthesis processes of legal professionals. By leveraging interactive dialogue among specialized agents, the system addresses a common limitation of existing legal AI tools, which often depend solely on fragmented inputs provided by litigants and fail to replicate the iterative dynamics of real consultations.

A distinctive feature of our work lies in the use of **real-world data**. Instead of relying on synthetic or simulated examples, we ground the framework in authentic legal aid case records, ensuring that the system reflects the fragmented narratives, incomplete facts, and practical constraints of real legal practice. This choice not only enhances empirical validity but also strengthens the relevance of our findings for deployment in real environments.

By integrating multi-agent interaction with real-world legal aid data, this study introduces a framework that better reflects the iterative nature of legal consultations and establishes a foundation for future evaluation and deployment.

## 2. Literature Review

### *2.1. Multi-Agent AI Systems*

Multi-agent systems (MAS) have long been studied in artificial intelligence for their capacity to model distributed reasoning, negotiation, and decision-making [14,15]. In the legal domain, early work focused on computational models of normative reasoning [16] and agent-based simulations of law-making or dispute resolution. More recent studies have explored the coordination of multiple agents to improve task specialization, robustness, and explainability [18]. Despite these advances, most MAS research in AI and law remains conceptual or simulation-based, with limited application to real-world legal service contexts.

### *2.2. Legal Chatbots and Conversational Agents*

Conversational agents for legal tasks have received increasing attention, particularly in the form of question-answering systems, contract assistants, and domain-specific chatbots [20]. These systems are often built on natural language processing pipelines [21] or, more recently, on large language models (LLMs) [22]. While such tools can provide quick responses to legal questions, they typically operate in a single-turn setting, generating answers directly from a user's initial input. As a result, they lack the capacity for multi-turn interrogation and clarification, which is essential for accurate fact collection. Recent work on user simulation for dialogue systems [23] has also inspired our design of the **LitigantTwin**, emphasizing controllable factual constraints and progressive information disclosure.

### *2.3. Our Propose: Legal Multi-agent System*

These research gaps motivate our proposal: a multi-agent framework for legal aid that leverages real-world case transcripts. The system replicates the iterative nature of lawyer–client consultations through structured questioning and constrained client modeling, and further integrates narrative generation with an automated evaluation module. By combining authentic data with architectural innovation, our approach directly addresses these limitations in prior work.

## 3. Overview of Multi-agent System

Building on the above design principles, this study introduces a multi-agent framework that simulates real-world legal consultations through interactive dialogue, progressively collecting and verifying case facts. The **LitigantTwin** functions as a constrained client proxy, restricted to factual units to prevent hallucinations and maintain role consistency. The **LegalAider** acts as a lawyer's assistant, applying structured questioning strategies and supported by a *Conflict Checker* and *Closer* to ensure efficiency and professionalism. The **NarrativeAgent** and **Evaluator** together form an output-and-verification loop: the former generates legally coherent narrative summaries safeguarded by checklists and strict modes, while the latter assesses them with similarity scoring, omission detection,

and calibrated alignment to human judgment. Finally, the **AutoPlaybook** enhances questioning through rule-based classification, LLM-assisted fallback, and dynamic switching, ensuring both specialization and robustness across case types.

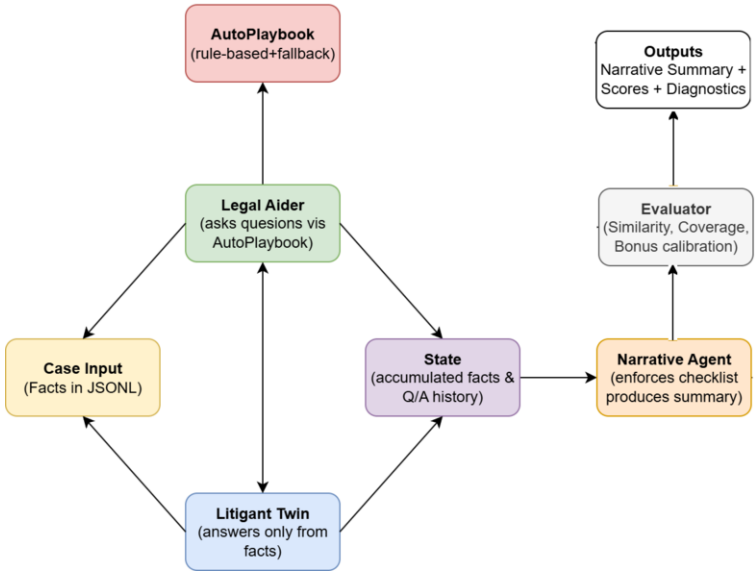



**Figure 1.** Overview of the Multi-Agent Framework

### *3.1. Litigant Twin*

Within the multi-agent framework, the **LitigantTwin** module draws inspiration from the **digital twin** concept but adapts it to a legal and conversational context rather than a physical or sensor-based environment. The objective is not to replicate a complete digital counterpart, but to maintain a fact-constrained virtual persona that mirrors how real litigants gradually disclose information during legal consultations. In this design, the system constructs a virtual "litigant persona" whose responses are strictly constrained by the **facts** field in the input dataset. The **LitigantTwin** does not generate information beyond what is explicitly provided, ensuring that simulated lawyer–client dialogues remain faithful to the original case record.

Technically, the module first segments the *facts* into minimal factual units, separated by punctuation (periods, question marks, exclamation points). These fact units are stored in a pool named *fact_chunks*, which serves as the exclusive source of information. The system also maintains a *revealed* list to track which facts have already been disclosed during the dialogue. When preparing a response, the LitigantTwin selects up to three unrevealed units and concatenates them into a "fact window," thereby balancing realism and interactivity: responses are factually accurate without prematurely exhausting all available information.

This design prevents hallucinations and maintains role consistency, producing outputs in a natural client-like voice rather than technical or legal jargon. The module also tracks which facts have already been revealed, releasing information incrementally to

mimic the gradual disclosure typical of real consultations. When a question cannot be answered based on available facts, the LitigantTwin outputs standardized responses such as “not sure” or “unclear,” preserving both authenticity and reliability.

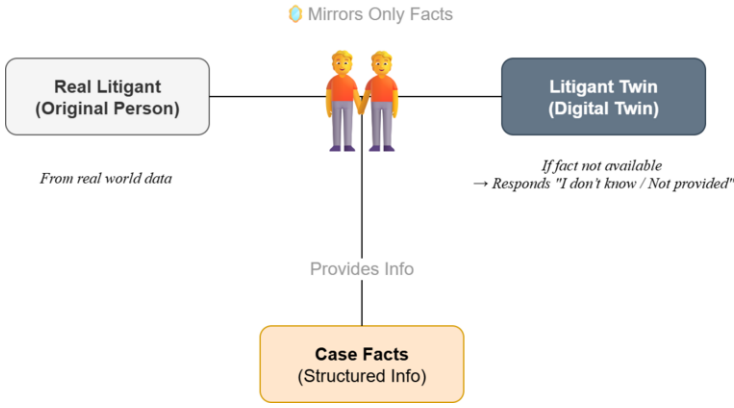



**Figure 2.** LitigantTwin as a digital mirror of facts. Responses are restricted to factual fragments, avoiding hallucination and preserving the client’s voice.

In sum, the LitigantTwin ensures **information source control** and **role consistency**, enabling the simulated client to respond only on the basis of authentic facts, while preserving traceability and verifiability of the dialogue. It should be noted that the **LitigantTwin** is not a digital twin in the strict technical sense. Rather, it is designed to emulate the behavioral realism of actual litigants, who often provide incomplete or selectively favorable information during consultations. Accordingly, the agent is allowed to respond with “not sure” or “unclear” when the required information is missing, reflecting authentic human uncertainty. Because the subsequent **Evaluator** module compares generated dialogues against verified ground-truth records, the **LitigantTwin** must not be permitted unrestricted generation or speculation. This constraint ensures factual integrity and preserves the reliability of downstream evaluation.

### *3.2. LegalAider*

The LegalAider functions as a lawyer’s assistant, orchestrating questioning, clarification, and case closing. It encapsulates the orchestration logic and defines a unified interface for the multi-turn dialogue process.

Upon initialization, the LegalAider loads a designated **playbook** corresponding to the case domain (e.g., consumer disputes, contract litigation, or family matters). The Orchestrator initiates interaction with an opening greeting, after which the LegalAider dynamically generates responses through the `reply()` method. Depending on context, it either asks further clarifying questions to gather facts or produces a case-closing summary when sufficient information has been collected.

The system enforces strict behavioral guidelines via system prompts, such as **Avoid redundancy, Focus on the client’s primary claim, Progress sequentially, Single focus per turn** The questioning process follows sequential and non-redundant principles, such as **Avoid redundancy, Focus on the client’s primary claim, Progress sequentially, Single focus per turn** to ensure clarity and structured JSON outputs.

To maintain robustness, the LegalAider includes a **fallback mechanism**: if the model fails to produce valid JSON, the system wraps the output into a minimal structure {"next_question": "..." } to ensure continuity. Additional modules enhance reliability: a *conflict checker* detects contradictions or inconsistencies and prompts clarifications, while a *closer* assembles the case summary under termination conditions. The closer is strictly limited to information surfaced in the dialogue and state, marking unknown elements explicitly as "unknown" rather than speculating.

Overall, the LegalAider's contribution lies not only in asking questions but in **how questions are asked**. Through carefully designed questioning principles, structured outputs, and conflict management, the LegalAider ensures logical progression of the interview, consistency of information, and resistance to hallucination.

### *3.3. Narrative Summary and Automated Evaluation*

At the conclusion of the dialogue, the system generates a comprehensive **narrative summary** of the collected facts, a task handled by the **NarrativeAgent** (*narrative_agent.py*).

The NarrativeAgent also supports two operational modes: (1) **strict mode**, relying solely on dialogue state without accessing ground-truth facts, and (2) **hybrid mode**, where ground facts may serve as a reference for evaluation but not be directly copied into the summary.

Evaluation is conducted by the **Evaluator**, which employs an LLM to perform both quantitative and qualitative assessments. It compares narrative summaries against the ground-truth facts, outputting (a) a similarity score between 0–1 and (b) diagnostic information, including missing spans, paraphrase pairs, and potential contradictions. Scoring rules are tiered: minor omissions allow scores up to 0.85; contradictions cap scores at 0.70; cases with both omissions and contradictions are limited to 0.60.

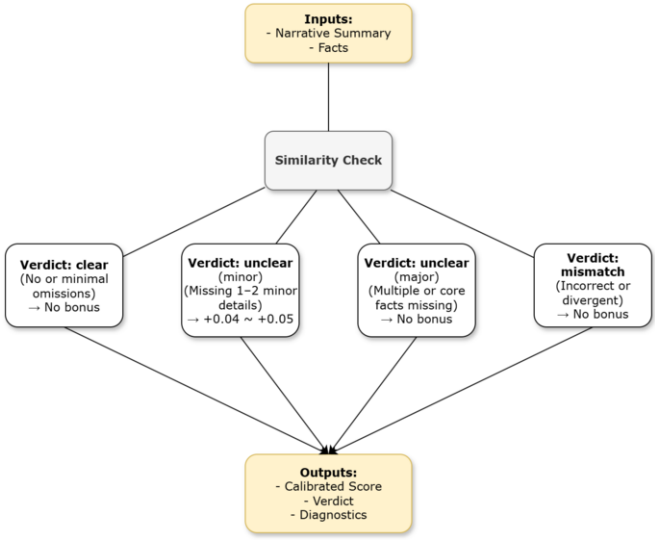



**Figure 3.** Evaluation workflow. The Evaluator compares AI-generated narratives against ground-truth facts, producing similarity scores and diagnostic outputs for omissions, paraphrases, and contradictions.

Together, the NarrativeAgent and Evaluator form the **output and verification layer** of the system. The NarrativeAgent ensures narrative completeness, while the Evaluator provides structured quality assessment. This dual-layer design enables scalable quantitative evaluation of large datasets while retaining fine-grained interpretability, addressing the challenge of balancing automation with legal rigor.

### *3.4. AutoPlaybook*

Because prompts can never be perfectly comprehensive, overly long or detailed instructions may actually reduce the LLM's ability to follow rules consistently. To address this limitation, in addition to the three core agents, we designed an **AutoPlaybook** module to support the questioning process of the *LegalAider*.

The design is inspired by how lawyers adopt different questioning strategies depending on the type of case. For example, in consumer disputes the emphasis is often placed on payment and contract clauses, while in family cases the focus is on kinship relations and procedural issues. Accordingly, the system includes multiple specialized *playbooks*, each containing prioritized dimensions of inquiry and corresponding key questions.

Technically, the AutoPlaybook first applies a set of rule-based classifiers, using regular expression matching across case titles, facts, and problem descriptions to score candidate playbooks. For instance, if the text contains terms such as "refund" or "consumer protection committee," the consumer playbook score increases significantly. If terms such as "divorce" or "custody" appear, the family case playbook is prioritized. Negative hints are also built in to prevent misclassification, such as excluding the "contract" category when the word appears in a family dispute. The system then selects the category with the highest score and loads the corresponding playbook.

If rule-based scoring is insufficient to produce a clear result, the AutoPlaybook invokes an LLM-based classifier to select the most appropriate category from a predefined list. When the LLMs' output still fails to meet expectations, the system defaults to a *common fallback* playbook to ensure that the process is not interrupted. This layered design allows the system to maintain dialogue continuity even in cases of misclassification.

Together, these modules constitute a closed-loop framework that balances professional orientation with robustness, enabling reliable and realistic multi-turn legal consultations.

## 4. Real-World Legal Data

The dataset used in this study consists of **real-world legal aid case transcripts and case summaries**, rather than artificially constructed or simulated synthetic data. These records were originally collected by the Legal Aid Society at National Taiwan University during actual consultation processes, including the litigant's narrative, lawyer's questions, and basic case registration information. For research purposes, all data underwent strict **de-identification**, removing names, phone numbers, addresses, and company identifiers, which were replaced with neutral placeholders (e.g., "Party A," "Party B," "Company X"), ensuring the protection of personal privacy.

In total, 100 legal-aid case records were compiled, covering the following common categories of legal aid cases:

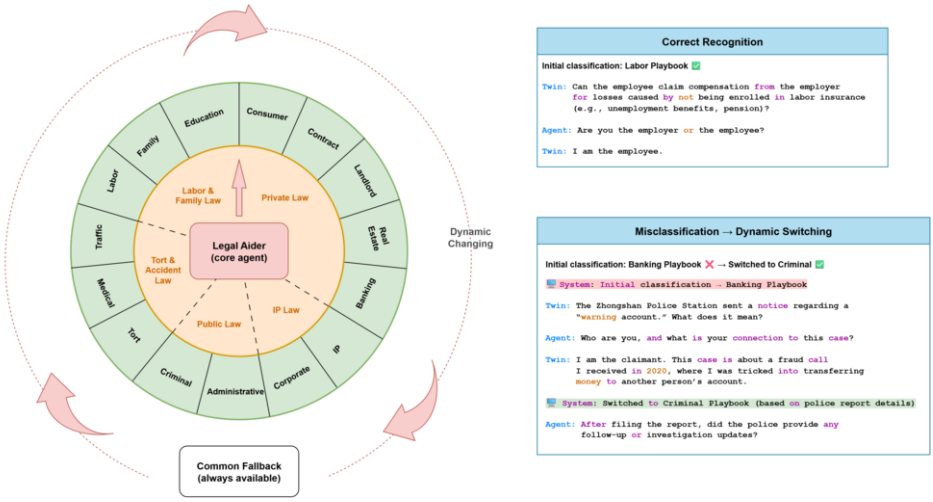



**Figure 4.** AutoPlaybook module. Balancing specialization and robustness, it dynamically selects or switches playbooks to ensure reliable and context-sensitive questioning.

1. **Consumer disputes:** e.g., sales contract conflicts, deposit refund requests.
2. **Family cases:** e.g., marriage, custody, inheritance.
3. **Debt and labor disputes:** e.g., loan repayment, wage disputes.
4. **Traffic accidents and criminal cases.**
5. **Property disputes:** e.g., ownership and usage rights.
6. **Tort liability and compensation.**

These heterogeneous case types allow us to test the adaptability of the system across different factual structures. And also, to facilitate multi-agent processing, the dataset was converted into **JSONL (JSON Lines)** format. This format is both human-readable and machine-parsable, supporting automated processing and reproducible experiments.

## 5. Experiment and Its Evaluation

### *5.1. Evaluation Indicators*

To comprehensively assess the performance of the proposed multi-agent system in legal fact collection and narrative generation, this study adopts a **three-tier evaluation framework** consisting of *Similarity*, and *Calibration*.

*A. Similarity (primary indicator).* Similarity is the principal metric for evaluating the semantic consistency and factual correctness between the generated narrative and the original facts. Using a large language model (LLM) as an evaluator, the system produces:Raw Score (ranging from 0.0 to 1.0), Verdict (*clear* / *unclear* / *mismatch*), Missing Spans. It allows semantically equivalent expressions to be recognized as correct even when not literally identical to the source text. Similarity allows semantically equivalent expressions to be recognized as correct even when not literally identical to the source text. This approach aligns more closely with human reasoning by assessing whether the LegalAider's narrative conveys the same case meaning as the factual record.

*B. Calibration (bonus adjustment rule).* During early experiments, we found that the Similarity metric was overly strict: minor omissions—such as missing dates or company names—often resulted in disproportionately low scores. To address this issue, a *bonus calibration rule* was introduced to moderate the penalty for omissions of secondary details. In the final evaluation, both the *Raw Score* and the *Calibrated Score* are retained, with the latter serving as the core reference for subsequent analyses.

### *5.2. Preliminary Results*

To validate the effectiveness of the proposed multi-agent framework, a series of simulation experiments were conducted using real-world legal aid cases. All experiments were performed under a unified environment, with **OpenAI GPT-4o mini** serving as the base language model. The system was configured through **few-shot prompting** and task-specific system instructions to ensure consistency across agents. Temperature settings were adjusted according to the functional characteristics of each agent: the **LitigantTwin** was assigned a temperature of 0.7 to simulate natural and conversational responses; both the **LegalAider** and the **NarrativeAgent** operated at a temperature of 0.2 to maintain stability and coherence in questioning and summarization; and the **Evaluator** ran at zero temperature to minimize randomness and enhance reproducibility.

In the overall results, the system achieved an average **Raw Score** of approximately 0.62, which increased to 0.67 after applying the calibration rule. This improvement indicates that the generated summaries exhibit stable semantic consistency with the original case facts. The average adjustment from calibration was around 0.04, with more than half of the cases receiving a positive bonus. This pattern suggests that although minor details were sometimes omitted, the overall event descriptions remained aligned with the core factual meaning. Table 1 summarizes the bonus effect across verdict types, illustrating how the calibration rule increased mean scores for *same event* and *unclear* cases while preserving semantic consistency.

**Table 1.** Bonus Effect by Verdict

| Verdict | Raw Score Mean | Calibrated Score Mean | Average Bonus |
|---|---|---|---|
| *different_event* | 0.23 | 0.30 | None |
| *same_event* | 0.80 | 0.85 | 0.05 |
| *unclear* | 0.60 | 0.63 | 0.03 |

This observation suggests that verdict classification alone is insufficient to capture the system's performance; instead, both score ranges and narrative completeness should be considered. Overall, the quantitative results demonstrate that the proposed framework maintains a medium-to-high level of semantic consistency across diverse case types. These findings further justify the inclusion of the calibration mechanism, which aligns automated evaluation outcomes more closely with human judgment.

## 6. Discussion and Conclusion

### *6.1. Strengths*

Unlike single-model approaches that generate summaries solely from initial inputs, our framework allows the **LegalAider** to probe with successive clarifying questions while

the **LitigantTwin** provides constrained responses. This gradual process fills missing details step by step, building a fuller and more accurate case narrative. Another strength lies in the **NarrativeAgent**, which transforms fragmented responses into narrative-style summaries tailored to legal documentation. By enforcing checklists and operating under a "strict mode," it ensures that essential elements such as dates, amounts, actions, and procedures are consistently included. Such narrative outputs are especially valuable in legal aid, where summaries can be directly reused for mediation or litigation without extensive manual restructuring.

Equally important is the role of the **Evaluator**, which introduces a systematic mechanism for automated quality control. Beyond assigning scores, it highlights omissions and contradictions, enabling both researchers and legal professionals to pinpoint weaknesses efficiently. The inclusion of a **bonus calibration** further reduces overly strict penalties, producing results that align more closely with human judgment. In addition, the **AutoPlaybook** balances domain specialization with robustness. By automatically identifying case types and selecting tailored questioning strategies, it helps the LegalAider focus on core disputes more effectively. When initial classification proves inaccurate, the system smoothly reverts to general rules, maintaining continuity and avoiding breakdowns in interaction.

Taken together, these design choices give the framework its distinctive advantages: structured interaction, narrative-oriented output, explainable evaluation, and strategy-driven questioning.

### *6.2. Future Work: Public Value and Legal Application*

This study highlights not only technical contributions but also significant public value. In legal aid consultations, litigants often present fragmented or unclear narratives, creating **information asymmetry**. By structuring multi-agent dialogue, the framework can consolidate scattered details into coherent accounts, reducing lawyer workloads and improving both efficiency and clarity. The **narrative summaries** generated are especially valuable, as they align with legal documentation requirements and can be used directly for mediation or court preparation, minimizing risks of omission or misinterpretation.

Future development should focus on refining evaluation methods, expanding data coverage, enhancing legal reasoning, and validating usability in real practice. Current metrics such as Similarity provide useful baselines but remain limited: the former only verifies mentions, while the latter relies heavily on LLM judgment. More robust approaches could integrate rule-based checks, semantic embeddings, and human–AI joint evaluation. Dataset expansion is also essential, extending beyond the current 100 cases in consumer, family, and labor disputes to include domains such as criminal, intellectual property, and environmental law, as well as cross-jurisdictional data.

Finally, enhancing **legal reasoning capabilities** and conducting **user-centered validation** are crucial. A dedicated reasoning agent could connect structured facts to statutes, producing preliminary opinions and completing the pipeline from fact collection to legal advice. At the same time, large-scale field studies involving lawyers, legal aid volunteers, and end-users will provide practice-oriented feedback for refinement. As a limitation, the current evaluation still relies primarily on automated comparison, we will seek confirmation from human experts to validate the system's outputs. This ensures that the current focus remains on the conceptual and practical foundations of the framework.

## References


[1] Jenkins J. What can information technology do for law. Harvard J Law Technol. 2008;21:589.

[2] Waltl B, Zec M, Matthes F. A data science environment for legal texts. In: Proceedings of the International Conference on Artificial Intelligence and Law (ICAIL). 2015. p. 193-194.

[3] Conrad JG, Al-Kofahi K, Zhao Y, Karypis G. Effective document clustering for large heterogeneous law firm collections. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2005. p. 177-187.

[4] Pacheco HR, Pavez MM. Contemporary challenges in legal informatics: Workshop INJU. In: 2016 11th Iberian Conference on Information Systems and Technologies (CISTI). 2016. p. 1-2.

[5] Lachmayer F, Cyras V. Visualization of legal informatics. J Vis Law. 2021;3:3-10.

[6] Sharma S, Gamoura S, Prasad DM, Aneja A. Emerging legal informatics towards legal innovation: Current status and future challenges and opportunities. Legal Inf Manage. 2021;21:218-235.

[7] Katz D, Dolin R. Legal informatics. Cambridge: Cambridge University Press; 2021.

[8] Šavelka J, Ashley KD. The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts. Front Artif Intell. 2023;6:1279794. doi:10.3389/frai.2023.1279794.

[9] Sun Z. A short survey of viewing large language models in legal aspect. ArXiv. 2023;abs/2303.09136. doi:10.48550/arXiv.2303.09136.

[10] Shaghaghian S, Feng L, Jafarpour B, Pogrebnyakov N. Customizing contextualized language models for legal document reviews. In: 2020 IEEE International Conference on Big Data (Big Data). 2020. p. 2139-2148. doi:10.1109/BigData50022.2020.9378201.

[11] Zhang D, Petrova A, Trautmann D, Schilder F. Unleashing the power of large language models for legal applications. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. 2023. doi:10.1145/3583780.3615993.

[12] Trozze A, Davies TP, Kleinberg B. Large language models in cryptocurrency securities cases: Can ChatGPT replace lawyers? ArXiv. 2023;abs/2308.06032. doi:10.48550/arXiv.2308.06032.

[13] Elwany E, Moore DA, Oberoi G. BERT goes to law school: Quantifying the competitive advantage of access to large legal corpora in contract understanding. ArXiv. 2019;abs/1911.00473.

[14] Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. *The Knowledge Engineering Review*, 10(2), 115–152.

[15] Stone, P., Veloso, M., & Tambe, M. (2000). Multiagent systems: A survey from a machine learning perspective. *Autonomous Robots*, 8(3), 345–383.

[16] Boella, G., van der Torre, L., & Verhagen, H. (2006). Introduction to normative multiagent systems. *Computational & Mathematical Organization Theory*, 12(2–3), 71–79.

[17] Liu, H.-W., Lin, C.-F., & Chen, Y.-J. (2019). Beyond State v Loomis: Artificial intelligence, government algorithmization and accountability. *International Journal of Law and Information Technology, 27*(2), 122–141. https://doi.org/10.1093/ijlit/eaz001

[18] Rahwan, I., & Simari, G. R. (Eds.). (2009). *Argumentation in Artificial Intelligence*. Springer.

[19] Susskind, R. (2023). *Tomorrow's Lawyers: An Introduction to Your Future* (2nd ed.). Oxford University Press.

[20] Gromova, E. A., Ferreira, D. B., & Begishev, I. R. (2023). ChatGPT and other intelligent chatbots: Legal, ethical and dispute resolution concerns. *Revista Brasileira de Alternative Dispute Resolution (RBADR), 5*(10), 153–175.

[21] Ashley, K. D. (2017). *Artificial Intelligence and Legal Analytics*. Cambridge University Press.

[22] Zheng, L., Guha, N., Anderson, B., Henderson, P., Ho, D. E., Nair, S., ... & Zhang, D. (2023). LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. *arXiv preprint* arXiv:2308.11462. https://arxiv.org/abs/2308.11462

[23] Zhao Y., Chen L., Liu H., Zhang D. User Simulation for Multi-Turn Dialogue Evaluation. *arXiv preprint arXiv:2402.13374*, 2024.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*

*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251592*

# Do LLMs Truly "Understand" when a Precedent Is Overruled?

Li ZHANG [a,1], Jaromir SAVELKA [b], and Kevin ASHLEY [a]
[a] *University of Pittsburgh*
[b] *Carnegie Mellon University*
ORCiD ID: Li Zhang https://orcid.org/0000-0003-0375-1793, Jaromir Savelka https://orcid.org/0000-0002-3674-5456, Kevin Ashley https://orcid.org/0000-0002-5535-0759

**Abstract.** Large language models (LLMs) with extended context windows show promise for complex legal reasoning tasks, yet their ability to understand long legal documents remains insufficiently evaluated. Developing long-context benchmarks that capture realistic, high-stakes tasks remains a significant challenge in the field, as most existing evaluations rely on simplified synthetic tasks that fail to represent the complexity of real-world document understanding. Overruling relationships are foundational to common-law doctrine and commonly found in judicial opinions. They provide a focused and important testbed for long-document legal understanding that closely resembles what legal professionals actually do. We present an assessment of state-of-the-art LLMs on identifying overruling relationships from U.S. Supreme Court cases using a dataset of 236 case pairs. Our evaluation reveals three critical limitations: (1) *era sensitivity* – the models show degraded performance on historical cases compared to modern ones, revealing fundamental temporal bias in their training; (2) *shallow reasoning* – models rely on shallow logical heuristics rather than deep legal comprehension; and (3) *context-dependent reasoning failures* – models produce temporally impossible relationships in complex open-ended tasks despite maintaining basic temporal awareness in simple contexts. Our work contributes a benchmark that addresses the critical gap in realistic long-context evaluation, providing an environment that mirrors the complexity and stakes of actual legal reasoning tasks.

**Keywords.** large language models, legal reasoning, overruled precedents, information retrieval, trustworthy AI

## 1. Introduction

The recent expansion of the context windows of large language models (LLMs) has opened up possibilities for their application in domains that require understanding of long, complex documents [16]. Transformer-based models [30] have evolved from the original 512-token limit to models capable of processing millions of tokens [15], with key developments including sparse attention mechanisms [11,28] and memory-efficient

[1] Corresponding Author: Li Zhang, liz239@pitt.edu
The full dataset can be accessed at https://github.com/lizhang-AIandLaw/Do-LLMs-Truly-Understand-When-a-Precedent-Is-Overruled

attention algorithms [15]. Alternative approaches have been proposed to extend long-context modeling without changing the model architecture, such as retrieval-augmented generation (RAG) [26,7] and the use of external memory [31].

However, developing effective benchmarks for evaluating long-context understanding remains a significant challenge in the field. Evaluations relying on synthetic tasks, simplified question-answering formats, or artificially constructed scenarios fail to capture the complexity and stakes of real-world document understanding tasks. The legal field, with its reliance on lengthy case files and intricate statutes [23], presents an ideal domain. This paper presents an investigation into the capabilities of long-context LLMs in a particularly consequential legal task: the identification of overruling relationships. In U.S. law, overruling occurs when a later court declares that a prior decision should no longer be followed as binding precedent; we adopt this as the operational definition for this study.

We introduce a new dataset of 236 U.S. Supreme Court case pairs from 1789 to 2025, each consisting of an overruling case and the case(s) it overruled. Using this dataset, we evaluate a selection of long-context LLMs on three distinct tasks designed to probe the limits of their comprehension. We provide the models with the full texts of the overruling case which explicitly states that it overruled the specific case(s), forcing them to navigate a long context to identify the overruling relationship between them.

This task is challenging even when judicial opinions explicitly use terms like "overruled." The underlying reasoning process is distributed across pages of text, requiring readers to integrate complex legal arguments across sections into a coherent logical chain. Judicial opinions are filled with specialized legal terminology and subtle nuances that require deep contextual understanding [1]. Our work addresses the critical gap in long-context evaluation by providing a benchmark that captures the realistic complexity of legal reasoning while maintaining the scientific rigor needed for meaningful model assessment.

## 2. Related Work

### *2.1. Long-Context Language Model Evaluation*

The evaluation of long-context models has led to the development of specialized benchmarks. LongBench [2] introduced a comprehensive evaluation suite covering multiple languages and domains, while LongEval [13] and GSM-Infinite [36] focused specifically on the degradation of performance as context length increases. The SCROLLS benchmark [25] provided domain-specific evaluation across various long-document tasks. LongDocURL [5] and MMLONGBENCH-DOC [17] built multimodal long document benchmarks integrating understanding and reasoning. These benchmarks have revealed that performance often degrades significantly when information is embedded deep within long contexts, particularly for tasks requiring complex reasoning rather than simple information retrieval.

Our work extends this line of research by introducing a novel legal reasoning task that specifically tests models' ability to identify overruling relationships across lengthy judicial opinions. Unlike previous benchmarks that focus on general document understanding, our task requires temporal reasoning, hierarchical relationship identification, and the integration of complex legal arguments distributed across extended text.

### 2.2. LLMs Application and Evaluation in Legal Domain

The application of LLMs in the legal domain has seen significant growth, driven by the promise of automating complex legal tasks [24,9] and improving access to legal information [27]. Several legal evaluation benchmarks have emerged to assess LLM performance in legal tasks. The LexGLUE benchmark [3] provides a unified evaluation framework covering multiple legal NLP tasks including case outcome prediction, legal text classification, and legal question answering. The CUAD dataset [12] focuses specifically on contract understanding and analysis, testing models' ability to extract and reason about complex contractual terms. Other benchmarks [10,34,18,21] offer evaluation methods covering reasoning, information extraction, and document analysis tasks across various legal domains. These benchmarks have demonstrated that while LLMs show promising results on some legal tasks, they struggle with complex reasoning that requires deep understanding of legal principles.

Our work contributes to this body of research by introducing novel tasks that specifically test LLMs' ability to identify overruling relationships. Previous evaluations also mainly focus on multiple choice QA and classification tasks while our study tests the models' ability to navigate real-world full-context judicial opinions by designing three tasks with a mixture of open-ended and closed-ended questions.

## 3. Dataset

To evaluate the performance of LLMs on the tasks of identifying overruled precedents, we created a new dataset of 236 U.S. Supreme Court case pairs in the period from 1789 to 2025. Each pair consists of a case that was explicitly overruled and the subsequent decision(s) that overruled it.

The foundation of our dataset is a resource maintained by the Constitution Annotated [4] consisting of 236 U.S. Supreme Court case pairs. The compilation process involved a comprehensive search of the Lexis database for all Supreme Court decisions containing the word "overrule" in the headnotes, syllabus, or opinion. A decision is included only when a majority of the Supreme Court has explicitly stated that it is being overruled. This strict criterion avoids the ambiguity inherent in legal commentary and ensures that each case in our dataset has been definitively identified as overruled by the Court itself. Conversely, cases that are merely distinguished, limited, or discredited without an explicit overruling are excluded. This conservative approach ensures the dataset's reliability. The results were then manually reviewed to confirm the Court's intent and to ensure that each case met the strict inclusion criteria.

For each overruling case of the 236 case pairs, we retrieved the full opinion texts from CourtListener [29], including majority and lead opinions as well as combined, concurring, and dissenting opinions. We then cleaned the raw downloads to retain only the textual content and conducted a manual review to verify content quality and completeness. The resulting curated corpus constitutes the dataset used in this study.

Figure 1 illustrates the distribution of word counts across our dataset. The 236 overruling cases range from the shortest case with 705 words to the longest with 116,455 words. The median case length of 17,607 words and mean case length of 19,368 suggest that most cases in our dataset require efforts to process and understand, making them ideal for testing the limits of LLM long-context comprehension.

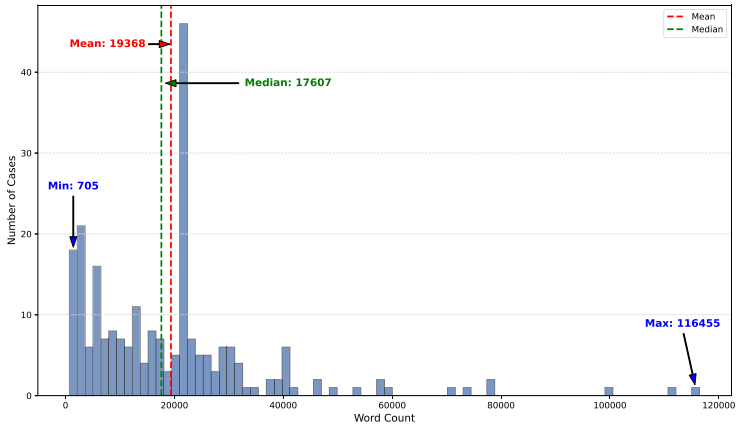



**Figure 1.** Word Count Distribution of 236 Overruling Cases in Our Dataset.

## 4. Experiment Design

To probe the capabilities of LLMs' long-context understanding in the legal domain, we designed three tasks centered on identifying overruling relationships.

### *4.1. Task 1: Open-Ended Identification*

The first task poses an open-ended question to LLMs: "Which case(s) was overruled by [the overruling case, e.g., "*Hohn v. United States*, 524 U.S. 236 (1998)"]? Here is the related context [case text of the overruling case]." In this prompt, we use the overruling case as the context and ask the LLM to identify the case(s) that were overruled by the overruling case.

*Rationale* This task simulates a realistic legal research query. To evaluate the open-ended responses, we employed `GPT-4o` as an automated judge, using a few-shot prompt to ensure consistent and accurate assessment against the ground truth. The core prompt was designed as "Do [the model's response, e.g., "*House v. Mayo*"] and the ground truth [the overruled case(s), e.g., "*House v. Mayo*, 324 U.S. 42 (1945)"] refer to the same case?"

The development of the few-shot examples involved two rounds of iterative testing. First, we compared GPT-4o's zero-shot judgment results against our ground truth annotations and analyzed the discrepancies to identify patterns in misclassification. Based on these findings, we refined the prompt by incorporating carefully selected examples that addressed the common failure modes. In the second round, we conducted a validation experiment where 10% of the responses were randomly selected for human expert review, demonstrating that GPT-4o's accuracy as a judge under the few-shot prompting approach was 98% consistent with human expert assessments.

### *4.2. Task 2: Closed-Ended Verification*

The second task is a verification query: "Did [the overruling case, e.g., "*Hohn v. United States*, 524 U.S. 236 (1998)"] overrule [the overruled case, e.g., " *House v. Mayo*, 324

U.S. 42 (1945)"]? Here is the related context [case text of the overruling case]." With this prompt, the model must answer "true," "false," or "unknown." The correct answer is always "true."

*Rationale* This task reduces the problem from open-ended QA to closed-ended verification. It allows us to isolate the model's ability to confirm or deny a specific legal relationship when presented with all necessary context, testing its comprehension and reasoning over long distances within the text.

### *4.3. Task 3: Reversed Closed-Ended Verification*

The third task presents a logically flawed question: "Did [the overruled case, e.g., "*House v. Mayo*, 324 U.S. 42 (1945)"] overrule [the overruling case, e.g., "*Hohn v. United States*, 524 U.S. 236 (1998)"]? Here is the related context [case text of the overruling case]." With this prompt, the model must answer "true," "false," or "unknown." The year of each case is provided in the context. The correct answer is always "false," as a precedent can only be overruled by a later decision.

*Rationale* This task serves as a control to test whether failures in the other tasks stem from a fundamental misunderstanding of temporal logic or from difficulties in understanding the long context. A model that understands chronology should consistently answer "false".

### *4.4. Models*

We evaluated five state-of-the-art LLMs known for their strong performance on long-context benchmarks. At the time of our experiments, the Gemini-Pro [8], Gemini-Flash, Qwen3 [32] represented the top 3 of the LongBench v2 leaderboard [2], and were selected to ensure our evaluation reflected the current state of the art. We also included GPT-5 [20] and Gemini-Flash-Lite to broaden our assessment. The models and their context window length (in tokens) are:

- `Qwen3-235B-A22B-Thinking-2507` (Qwen3) - Context Length: 262,144
- `Gemini-2.5-flash-lite-preview-06-17` (Gemini-Flash-Lite) - Context Length: 1,000,000
- `Gemini-2.5-flash-preview-05-20` (Gemini-Flash) - Context Length: 1,000,000
- `Gemini-2.5-pro-preview-06-05` (Gemini-Pro) - Context Length: 1,000,000
- `GPT-5-2025-08-07` (GPT-5) - Context Length: 128,000

### *4.5. Experimental Setup*

All experiments were conducted with a `temperature` of 0.1, a `top_p` of 0.5, and a structured json output format, to encourage deterministic responses. For GPT-5 model, we use the default parameters as set by OpenAI, as the relevant parameter controls are not publicly exposed.

## 5. Empirical Results

The performance of the selected LLMs is summarized in Table 1.

**Table 1.** Model Performance Across Tasks (Accuracy in %, higher is better; ± denotes 95% confidence intervals, $n = 236$).

| Model | Task 1 | Task 2 | Task 3 |
|---|---|---|---|
| Qwen3 | 2.12% ± 1.84% | 1.27% ± 1.42% | 97.46% ± 2.02% |
| Gemini-Flash-Lite | 0.85% ± 1.01% | 11.86% ± 4.08% | 91.95% ± 3.52% |
| Gemini-Flash | 33.47% ± 6.03% | 6.36% ± 3.12% | 62.71% ± 6.12% |
| Gemini-Pro | **73.31% ± 5.65%** | 42.37% ± 6.25% | **100% ± 0%** |
| GPT-5 | 71.19% ± 5.78% | **75.42% ± 5.47%** | **100% ± 0%** |

### *5.1. Task 1: Open-Ended Identification*

In the open-ended identification task, the models struggled to correctly identify the overruled case. The best-performing model, Gemini-Pro, achieved an accuracy of 73.31% (173 out of 236), closely followed by GPT-5 with 71.19% (168 out of 236). The other models performed worse. This suggests that even state-of-the-art models have difficulty with understanding the overruling relationship in a long legal context.

There are three patterns in the errors. First, when an overruling case overturns multiple precedents, models often extract only one of them, yielding a partially correct answer. Second, models sometimes confuse cases that are merely mentioned in the text with the case actually overruled—we refer to this as *confusion* errors, where models incorrectly identify a case that appears in the text but is not the overruled precedent. Third, models hallucinate a case that is not mentioned in the opinion at all—we define *hallucination* as the generation of case names or citations that do not exist in the provided context; for example, in Gemini-Flash-Lite's response to *Garland v. Washington*, the model produced *Miles v. Graham*, which is neither cited in the text nor overruled by *Garland*. In the calculation of the accuracy, we only count the cases where the model correctly identified all overruled cases. Table 2 provides illustrative examples of these patterns.

**Table 2.** Examples of Model Responses and Error Patterns for Task 1

| Overruling Case | Overruled Case | Model Response | Type |
|---|---|---|---|
| *Hohn v. United States*, 524 U.S. 236 (1998) | *House v. Mayo*, 324 U.S. 42 (1945) | *House v. Mayo*, 324 U.S. 42 (1945) | Correct |
| *Payne v. Tennessee*, 501 U.S. 808 (1991) | *South Carolina v. Gathers*, 490 U.S. 805 (1989) & *Booth v. Maryland*, 482 U.S. 496 (1987) | *South Carolina v. Gathers*, 490 U.S. 805 (1989) | Missing Case |
| *Garland v. Washington*, 232 U.S. 642 (1914) | *Crain v. United States*, 162 U.S. 625 (1896) | *Rogers v. Peck*, 199 U. S. 425, 435 (1905) | Confusion |
| *Garland v. Washington*, 232 U.S. 642 (1914) | *Crain v. United States*, 162 U.S. 625 (1896) | *Miles v. Graham*, 268 U.S. 501 (1925) | Hallucination |

### *5.2. Task 2 & 3: Closed-Ended Verification*

The closed-ended verification of Task 2 yielded more varied results. GPT-5 emerged as the top performer in this task, achieving an accuracy of 75.42% (178 out of 236), significantly outperforming Gemini-Pro which achieved 42.37% (100 out of 236). GPT-5 also showed a much lower rate of abstention, answering "unknown" in only 3 cases

compared to Gemini-Pro's 131 abstentions. However, this low abstention rate comes with a significant cost: GPT-5 incorrectly labeled 55 cases as "False," representing a substantial number of false negatives.

While GPT-5's low abstention rate might suggest higher confidence in its reasoning capabilities, it could also be interpreted as a tendency toward overconfident incorrectness. This trade-off highlights the need for more sophisticated uncertainty quantification mechanisms that can distinguish between genuine epistemic uncertainty and mere comprehension gaps. The other models showed a significant drop in performance, with Gemini-Flash answering "unknown" in the majority of cases. This suggests that even when the task is simplified to verification, most models struggle to confidently identify the correct legal relationship.

The reversed closed-ended verification of Task 3 showed a marked improvement in performance across all models. The accuracy was higher than in Task 2. GPT-5 and Gemini-Pro achieved high accuracy (100%), correctly identifying the logical impossibility in all cases. Qwen3 also performed well with 97.5% accuracy.

### *5.3. Discussion*

#### *5.3.1. Era Sensitivity*

To probe temporal robustness in the open-ended identification task (Task 1), we stratified cases by overruling case decision year (the earliest overruling case happened in 1810) and measured models' accuracy across the fixed historical intervals that reflect changes in case density and legal language evolution. There are 17 case pairs from 1810–1881, 65 case pairs from 1882–1953, and 154 case pairs from 1954–2025. This quantitative view in Table 3 aligns with our qualitative observations and clarifies where failures concentrate.

**Table 3.** Model Performance by Historical Intervals in Task 1 (Accuracy in %, higher is better).

| Model | 1810–1881 | 1882–1953 | 1954–2025 |
|---|---|---|---|
| Qwen3 | 0.00% | 1.54% | 2.60% |
| Gemini-Flash-Lite | 0.00% | 0.00% | 1.30% |
| Gemini-Flash | 17.65% | 24.62% | 38.96% |
| Gemini-Pro | **64.71%** | **72.31%** | 74.68% |
| GPT-5 | 35.29% | 53.85% | **82.47%** |

The temporal stratification reveals era sensitivity across all models, with performance degrading as we move backward in time. In the earliest period (1810–1881), the performance gap is stark: Qwen3 and Gemini-Flash-Lite achieve 0% accuracy, completely failing to identify any overruling relationships in this historical interval. The best-performing model in this period, Gemini-Pro, manages only 64.71% accuracy, while GPT-5 struggles at 35.29%. This performance drop in the earliest cases suggests that models trained primarily on modern legal texts lack the linguistic and conceptual frameworks needed to parse 19th-century judicial opinions, which often employ archaic legal terminology, different citation conventions, and distinct argumentative structures [19].

The middle period (1882–1953) shows a moderate improvement. Here, Gemini-Pro maintains its strong performance at 72.31% accuracy, demonstrating remarkable temporal consistency. GPT-5 shows significant improvement to 53.85%, while Gemini-Flash

reaches 24.62%. However, the weaker models continue to struggle: Qwen3 achieves only 1.54% accuracy, and Gemini-Flash-Lite reaches 0.00%. This intermediate period represents a transitional phase where legal language begins to modernize but still retains many historical characteristics that challenge contemporary models. The performance improvements suggest that models can adapt to some degree of historical variation, but their capabilities remain fundamentally limited when dealing with pre-modern legal discourse.

The modern period (1954–2025) represents the models' comfort zone, with significantly improved performance across the board. GPT-5 achieves its highest accuracy at 82.47%, significantly outperforming Gemini-Pro's 74.68%. Gemini-Flash shows improvement to 38.96% while remaining limited. This improvement in the modern era reflects several factors: the models' training data heavily emphasizes contemporary legal texts and the legal reasoning patterns align more closely with what the models have learned during training [6]. The fact that even the best-performing models achieve only around 80% accuracy underscores the fundamental challenge of legal reasoning tasks, while the precipitous drop in performance for earlier periods quantifies the temporal bias that underlies the open-ended reasoning failures we observed in our experiments.

### *5.3.2. Hesitation in Verification vs. Confidence in Refutation*

The high number of "unknown" responses in Task 2 is revealing, as shown in Figure 2. It suggests that most models are not simply guessing, but are actively assessing their own uncertainty. When the models are unable to reason and understand the case relationship embedded in the text, they are hesitant to make a definitive judgment. This is a desirable trait in a legal AI system, as it is preferable for a model to admit its own ignorance rather than to provide a confident but incorrect answer. However, the high rate of abstention in this task suggests that most models' threshold for certainty is set too high, preventing them from making correct judgments even when the evidence is strong. GPT-5's low abstention rate (1.27%, 3 out of 236) stands in stark contrast to this pattern, suggesting a different approach to uncertainty assessment.

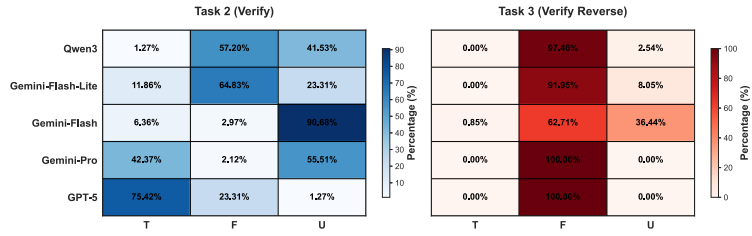



**Figure 2.** Confusion Matrix for Verification Tasks. T, F, and U stand for True, False, and Unknown respectively.

A contrast emerges when comparing the results of Task 2 and Task 3. In Task 2, when asked to verify a correct overruling relationship, the models frequently defaulted to "unknown," even when the answer was clearly stated in the text. In contrast, when presented with the temporally impossible proposition in Task 3, the same models answered with high confidence and accuracy, with GPT-5 and Gemini-Pro achieving 100% accuracy and Qwen3 achieving 97.46%.

This pattern of behavior reveals a fundamental limitation in the models' approach to legal reasoning. The contrast between Task 2 and Task 3 performance demonstrates that models rely on shallow logical heuristics rather than deep legal comprehension. In Task 3, models can successfully apply simple temporal logic—recognizing that an earlier case cannot overrule a later one—without actually understanding the complex legal reasoning embedded in the judicial opinions. This high performance is achieved through basic chronological awareness rather than genuine comprehension of legal arguments.

### *5.3.3. Temporal Reasoning*

Our analysis also reveals an interesting pattern in temporal reasoning across different task contexts. In the open-ended identification task (Task 1), we observed what we term "context-dependent temporal reasoning failures," as illustrated in Table 4. These failures show that models can create temporally impossible relationships, such as suggesting that a 1914 case overruled a 1925 case, even when the case names include the years. Such responses indicate a breakdown in temporal causality, where a case from an earlier year is incorrectly identified as overruling a case from a later year.

**Table 4.** Examples of Context-Dependent Temporal Reasoning Failures from Task 1. LLMs generated logically impossible scenarios where a subsequent case was overruled by a preceding one.

| **Case Relationship / Query** | **Year** |
|---|---|
| *Query Case (Overruling Case):* | |
| Garland v. Washington | 1914 |
| *Model's Response (Overruled Case):* | |
| Miles v. Graham | 1925 |

However, when we examine the results from Task 3 (the reversed verification task), a different picture emerges. As shown in Table 1, models demonstrate remarkably high accuracy in identifying temporally impossible propositions. GPT-5 and Gemini-Pro achieved 100% accuracy, while Qwen3 achieved 97.46% accuracy in correctly rejecting statements like "Did *Booth v. Maryland* (1987) overrule *Payne v. Tennessee* (1991)?"—a clear temporal impossibility.

This contrast suggests that the models are not fundamentally lacking in temporal reasoning capabilities or awareness of legal case chronology. Rather, the issue lies in the complexity of the reasoning required in different contexts. In Task 3, the models are presented with a straightforward logical contradiction that they can easily identify using basic temporal logic. In contrast, Task 1 requires more sophisticated reasoning—the models must understand complex legal relationships, extract relevant information from dense legal texts, and apply temporal logic in an open-ended context.

## 6. Limitations and Future Work

Our study has several limitations. First, our dataset focuses exclusively on U.S. Supreme Court cases, which may not generalize to other jurisdictions or legal systems with different precedent structures. Second, we only examined cases with explicit overruling language, excluding more nuanced legal developments like distinguishing, limiting, or discrediting precedents. Third, our evaluation relies on automated assessment using GPT-

4o, which, while validated against human experts, may introduce its own biases. Finally, we tested a limited set of models at a specific point in time, and rapid advances in LLM technology may render some findings obsolete.

Future work should address these limitations by expanding to international legal systems, incorporating more subtle legal relationships, and developing more sophisticated evaluation metrics. Additionally, research should explore architectural improvements for long-context understanding, such as legal domain-specific pre-training, better uncertainty quantification, and the potential of RAG systems [14,22] to enhance LLMs' long-context understanding capabilities. Exploring agent-based approaches [35,33] for enhancing LLMs' long-text capabilities could also lead to more consistent and reliable legal AI systems. The development of legal reasoning benchmarks that test deeper comprehension would further advance the field, along with more detailed quantitative analysis of error patterns across different models to strengthen qualitative insights into specific failure modes.

## 7. Conclusion

This paper presents an evaluation of state-of-the-art LLMs on the legal reasoning task of identifying overruling relationships from U.S. Supreme Court cases. Our investigation reveals three fundamental limitations that challenge the practical deployment of these models in legal contexts.

First, we demonstrate that models exhibit era sensitivity. The temporal stratification reveals era sensitivity across all models in our evaluation, with performance degrading as we move backward in time. This temporal bias reveals that extended context windows alone cannot overcome the fundamental challenge of understanding legal language across different historical periods, suggesting that models lack the linguistic and conceptual frameworks needed to parse judicial opinions from earlier eras that employ archaic legal terminology, different citation conventions, and distinct argumentative structures.

Second, our analysis reveals the models' reliance on shallow logical heuristics rather than deep legal comprehension. Models can successfully apply simple temporal logic—such as recognizing that an earlier case cannot overrule a later one—without actually understanding the complex legal reasoning embedded in the judicial opinions. The high performance on Task 3 is achieved through basic chronological awareness rather than genuine comprehension of legal arguments, highlighting a fundamental gap between surface-level logical reasoning and deep legal understanding.

Third, we identify context-dependent temporal reasoning failures that highlight the complexity of legal reasoning tasks. Models can correctly identify simple temporal contradictions but fail to maintain temporal consistency when processing complex, open-ended legal relationships. This finding indicates that the challenge lies not in basic logical reasoning, but in integrating temporal awareness with deep legal understanding under cognitive load.

The empirical results underscore a critical gap between current LLM capabilities and legal practice requirements. Even the best-performing model achieved only 73.31% accuracy on the core identification task, falling short of the reliability standards needed for legal applications. This performance gap is particularly concerning given that our dataset focuses on cases with explicit overruling language, representing a straightforward scenario for legal relationship identification.

## References


[1] Kevin D Ashley. *Artificial intelligence and legal analytics: new tools for law practice in the digital age*. Cambridge University Press, 2017.

[2] Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, et al. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *arXiv preprint arXiv:2412.15204*, 2024.

[3] Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. Lexglue: A benchmark dataset for legal language understanding in english. *arXiv preprint arXiv:2110.00976*, 2021.

[4] Congressional Research Service. Constitution of the united states: Analysis and interpretation. https://constitution.congress.gov.

[5] Chao Deng, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, et al. Longdocurl: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. *arXiv preprint arXiv:2412.18424*, 2024.

[6] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[7] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1), 2023.

[8] Google. Model cards. https://modelcards.withgoogle.com/model-cards, 2025. Accessed on 2025-08-26.

[9] Morgan Gray, Li Zhang, and Kevin D Ashley. Generating case-based legal arguments with llms. In *Proceedings of the 2025 Symposium on Computer Science and Law*, pages 160–168, 2025.

[10] Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in neural information processing systems*, 36:44123–44279, 2023.

[11] Chi Han, Qifan Wang, Hao Peng, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. Lm-infinite: Zero-shot extreme length generalization for large language models. *arXiv preprint arXiv:2308.16137*, 2023.

[12] Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268*, 2021.

[13] Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. Longeval: Guidelines for human evaluation of faithfulness in long-form summarization. *arXiv preprint arXiv:2301.13298*, 2023.

[14] Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. Retrieval augmented generation or long-context llms? a comprehensive study and hybrid approach. *arXiv preprint arXiv:2407.16833*, 2024.

[15] Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024.

[16] Jiaheng Liu, Dawei Zhu, Zhiqi Bai, Yancheng He, Huanxuan Liao, Haoran Que, Zekun Wang, Chenchen Zhang, Ge Zhang, Jiebin Zhang, et al. A comprehensive survey on long context language modeling. *arXiv preprint arXiv:2503.17407*, 2025.

[17] Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems*, 37:95963–96010, 2024.

[18] Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. Hallucination-free? assessing the reliability of leading ai legal research tools. *Journal of Empirical Legal Studies*, 22(2):216–242, 2025.

[19] David Mellinkoff. *The language of the law*. Wipf and Stock Publishers, 2004.

[20] OpenAI. Gpt-5 system card. https://cdn.openai.com/gpt-5-system-card.pdf, 2025. Accessed on 2025-08-26.

[21] Nicholas Pipitone and Ghita Houir Alami. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. *arXiv preprint arXiv:2408.10343*, 2024.

[22] Zehan Qi, Rongwu Xu, Zhijiang Guo, Cunxiang Wang, Hao Zhang, and Wei Xu. Long$^2$rag: Evaluating long-context & long-form retrieval-augmented generation with key point recall. *arXiv preprint arXiv:2410.23000*, 2024.

[23] John B Ruhl. Law's complexity: a primer. *Ga. St. UL Rev.*, 24:885, 2007.

[24] Jaromir Savelka and Kevin D Ashley. The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts. *Frontiers in Artificial Intelligence*, 6:1279794, 2023.

[25] Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, et al. Scrolls: Standardized comparison over long language sequences. *arXiv preprint arXiv:2201.03533*, 2022.

[26] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*, 2023.

[27] Francesco Sovrano, Kevin Ashley, Peter Leonid Brusilovsky, and Fabio Vitali. How to improve the explanatory power of an intelligent textbook: a case study in legal writing. *International Journal of Artificial Intelligence in Education*, pages 1–35, 2024.

[28] Jiaming Tang, Yilong Zhao, Kan Zhu, Guangxuan Xiao, Baris Kasikci, and Song Han. Quest: Query-aware sparsity for efficient long-context llm inference. *arXiv preprint arXiv:2406.10774*, 2024.

[29] The Free Law Project. RECAP archive, 2020.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[31] Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. *Advances in Neural Information Processing Systems*, 36:74530–74543, 2023.

[32] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[33] Li Zhang and Kevin D Ashley. Mitigating manipulation and enhancing persuasion: A reflective multi-agent approach for legal argument generation. *LCIC @ ICAIL 2025*, 2025.

[34] Li Zhang, Morgan Gray, Jaromir Savelka, and Kevin D Ashley. Measuring faithfulness and abstention: An automated pipeline for evaluating llm-generated 3-ply case-based legal arguments. *ASAIL @ ICAIL*, 2025.

[35] Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Arik. Chain of agents: Large language models collaborating on long-context tasks. *Advances in Neural Information Processing Systems*, 37:132208–132237, 2024.

[36] Yang Zhou, Hongyi Liu, Zhuoming Chen, Yuandong Tian, and Beidi Chen. Gsm-infinite: How do your llms behave over infinitely increasing context length and reasoning complexity? *arXiv preprint arXiv:2502.05252*, 2025.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*

*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251593*

# Multilingual Legislative Definitions Retrieval and Generation Using LLM and Agentic AI

Leonardo ZILLI, Michele CORAZZA, Monica PALMIRANI and Salvatore SAPIENZA

*CIRSFID - ALMA AI, University of Bologna, Italy*
ORCiD ID: Leonardo Zilli https://orcid.org/0009-0007-4127-4875, Michele Corazza https://orcid.org/0000-0002-7288-6635, Monica Palmirani https://orcid.org/0000-0002-8557-8084, Salvatore Sapienza https://orcid.org/0000-0002-5429-5217

**Abstract.** The application of AI-based methods and Large Language Models (LLMs) to the legislative domain poses unique challenges, including the extensive usage of normative references, the complexity of legal language, as well as the ever-changing nature of legal documents. We propose a multilingual (English-Italian) LLM-based method to both retrieve and generate legislative definitions in the context of the European and Italian legislation. These definitions are a crucial aspect of legislative documents, as they create new meaning for specific concepts, and their generation is an open challenge for any automatic method. New definitions should not conflict with pre-existing ones, be consistent with the specific legal domain (e.g., food, energy, finance), and instead leverage them when necessary. Our method fosters a Retrieval Augmented Generation approach, using LLM and Agentic AI, which considers the validity of existing definitions, the hierarchy of legal sources, and investigates strategies to mitigate hallucinations in the generation of definitions. We provide a quantitative and qualitative evaluation of the results of our experiments.

**Keywords.** Definitions, Legislation, Large Language Models, Agentic AI, RAG

## 1. Introduction

Drafting legislative documents is a fundamental component of the legislative process and a pillar of the normative system. This complex task is carried out by specialized legislative experts, and it is demanding even for experienced professionals. In this context, the integration of Artificial Intelligence (AI) tools could offer valuable support across a wide range of drafting-related activities.

Among such tasks, identifying the into force correct legislative definitions is a crucial step in drafting laws. In legal context, "Definitions" typically follow a pattern that establishes a clear relationship between a term (*definiendum*) and its description (*definiens*). This structure often takes the form: "*definiendum* means *definiens*" (e.g., *"domain" means one or several data sets that cover specific topics;"*). The process of drafting definitions is composed of three main tasks.

1. Identifying the appropriate definiendum according to the domain of the act and avoiding ambiguities with other similar terminology already used, with different meaning, in other context (e.g., *"data"* in data protection law has a certain meaning, but it has a different one in the Data Act).
2. Retrieving existing legislative definitions of the same term, to prevent overlapping and conflicts between different descriptions.
3. If no legislative definition of the definiendum exists, drafting an appropriate definiens for the term, bearing in mind "neighbor" definitions in the same domain and the necessity of avoiding ambiguity.

Given such complexity (*identification-retrieval-drafting*), this paper aims to identify possible methods of increasing the quality and the efficacy of legislative definition drafting, in particular by proposing a combined pipeline which simplifies the retrieval and drafting stages, which present both a higher degree of complexity and less discretion on the drafter, thus being prone to the development of AI-assisting tools.

While the progressive spread of Large Language Models (LLMs) and Agentic AI in the legal domain [1] allows new possibilities in document retrieval and text generation, legal theory demands significant adaptations of such computational approaches to the domain of law [2]. In particular, it is worth considering that legal texts bear a significant complexity in the handling of temporal aspects: repealed, no longer in force, or amended definitions should not be used in the generation of a new definition that should be applicable today. Finally, some parts of the text such as legal references (e.g., "Regulation 2016/679") have a specific meaning which shall be properly handled and evaluated not only syntactically, but also from a temporal or spatial/jurisdictional perspective.

While acknowledging such challenges, this paper seeks answers to the following questions:

- **RQ1**: to what extent LLMs and Agentic AI can contribute to the retrieval of existing legislative definitions from legal corpora?
- **RQ2**: to what extent LLMs and Agentic AI can contribute to the generation of new legislative definitions?
- **RQ3**: to what extent LLMs and Agentic AI can manage peculiar elements of the legal domain, namely jurisdiction, temporal validity, and normative references?

## 2. Background

In the application of AI to the legislative domain, a multitude of peculiar challenges arise from theoretical and practical issues that are intrinsically part of any legislative document [3].The scope of this study regards legislative definitions. They can have a multitude of characteristics and can be classified according to the following types:

- **Definitions as constitutive rules** introduce new concepts, principles or institutions and constitute the "poetic part" of the law. They act as a shortcut between a term and its legal meaning (e.g., "X counts as Y in C"). The majority of the definitions used in this study belong to the family of constitutive rules
- **Definitions as exceptions** support legal reasoning by specifying exceptions or derogations. These function as sub-definitions that extend, reduce, or exclude cases based on conditions. Definitions may broaden a category (e.g., "cabin luggage"), narrow its scope (e.g., restricting "placing on the market"), or explicitly exclude cases (e.g., excluding air-cushion vehicles from the term "ship").

- **Definitions as memberships** preserve consistency in legal concepts while maintaining adaptability over time. They typically follow an "is-a" structure, integrating common-sense or emerging societal terms into existing legal categories.
- **Definitions as equivalences** establish a semantic equivalence between two terms, often using formulations such as "means an offence equivalent to". These definitions frequently refer to external normative sources, to which they anchor their definition.
- **Definitions as interpretations** provide clarifications, specifications, or technical explanations without creating new legal categories. For example, defining the "length of a vessel" based on physical measurement standards.
- **Definitions as technical descriptions** present purely technical information without additional legal function. An example includes defining "a mile" as exactly 1852 meters.

The different nature of definitions is not the only critical issue. The structured nature of legislative documents, which are composed of chapters, sections, articles, lists with points, etc., is semantically relevant for the interconnection between provisions that determines the first criterion of legal reasoning of the legal text (e.g., the position of the points in the list is meaningful). Another issue that arises when dealing with legislative documents is the usage of normative references, meaning portions of text that are used to invoke information from another document. Their usage is pervasive in legislative definitions, allowing a single definition to be used across multiple laws (e.g., "personal data" means personal data as defined in point 1 of Article 4 of Regulation (EU) 2016/679). Finally, one issue arises when, over time, legislative documents are amended, repealed, thus affecting also their definitions. This means that each document can exist in multiple (consolidated) versions that were in force at different times, with different text for the same term of definition.

To mitigate the complexity of these challenges, across our experiments we leverage the Akoma Ntoso XML standard [4,5], a standard for legal documents which supports a variety of legal documents, including legislative ones. Akoma Ntoso allows our approach to be time-aware and to deal with documents using point-in-time consolidated versions. The Akoma Ntoso standard also supports the markup of structural elements of the document (article, paragraph, etc), the annotation of normative references which contain the URI of the referenced document, and extensive support for various types of metadata. In particular, AKN includes a sophisticated methodology for the annotation of all the types of definitions, especially those that are split into many provisions but semantically connected (e.g., definition splits in two article, in different points in the list).

## 3. Related Work and motivation

The prevalence of Natural Language Processing tools and LLMs in various multidisciplinary fields is one of the main trends of the last years. In the legal domain, this trend is also prevalent [6], since the automatic analysis of legal text can unlock a multitude of benefits for the legal community. Legal text pose unique challenges, and even state-of-the-art models struggle to produce satisfactory results [7]. Nevertheless, like in other fields, the prevalence of transformer-based models [8] first and LLMs later [9,10] has contributed to the field of legal informatics. Lately, applications based on Retrieval Augmented Generation (RAG) [11] can mitigate some of the shortcomings of LLMs in the

legal domain by grounding the models with accurate information.

This aspect is crucial for models that handle legislative documents, as these enter into force, are amended and abrogated continuously, meaning that models would be out of date in a very short amount of time. Interestingly, one of the application that involve RAG for the augmentation of LLMs is one where it was used to ground a model for question answering in the legal domain [12,13]. In a similar fashion, LLMs have been used to identify and map concepts in legislation with the goal of verifying EU harmonisation [14]. In the context of AI applications that handle legislative definitions, one relevant contribution regards the creation of a suggestion tool based on the European EuroVoc thesaurus[1] to suggest definition inside the LEOS editor [15]. Another approach, which is based on a generative LLM, is LexDrafter, which uses a RAG-based framework for the generation of European legislative definitions [16], enhancing the prompt with contextual information.

## 4. Datasets

Our experiments were performed on three datasets (in Italian and English, given the legal knowledge of the research group and by the legal expert evaluators), belonging to two different jurisdictions::

- **Eur-Lex**: a set of European Regulations and Directives, which were downloaded in FORMEX format from the eur-lex portal and converted in Akoma Ntoso XML. Out of the complete dataset, which contains 14305 documents from 2010 to 2021, 889 have been annotated with definitions obtained using a symbolic AI approach. They are in English.
- **Italian legislation in Normattiva**: a set of 3196 Akoma Ntoso XML documents extracted from Normattiva, the portal containing Italian laws. Out of those, 401 contain annotated normative references, which were extracted using linguistic patterns in the documents. They are in Italian.
- **Chamber of Deputies Bill (PDL)**: a set of 3709 Italian bills from the 18th and 19th legislatures, extracted by us from the Italian Chamber of Deputies and converted to Akoma Ntoso XML. Out of those, 78 contain legislative definitions. They are in Italian.

Throughout these documents, we used simple rule-based systems to annotate the legislative definitions contained within them, using the formulas adopted for definitions in the respective legislative documents (among others, "{term} means ..." in EU documents, "per {term} si intende" for Italian documents). The annotation was integrated in the Akoma Ntoso format using the appropriate elements ("def" for the definendum, "defBody" for the definiens). The documents are stored in an eXist XML database to allow the system to extract data and metadata from the documents leveraging Akoma Ntoso.

## 5. Methods

The system uses the LangGraph[2] framework to produce a conversational agent which is able to perform two tasks:

1. The retrieval of existing legislative definition in the two jurisdictions;

[1] https://eur-lex.europa.eu/browse/eurovoc.html
[2] https://github.com/langchain-ai/langgraph

2. If no definition is found, the system generates a new one.

In order to perform these operations, we use an agent-based approach, where the model itself is able to select and use tools to achieve its goals. The first step in both pipelines is the extraction of a series of parameters, which are provided in plain text by the user and extracted by the model:

- **User query:** the entire text provided by the user to the model;
- **Definiendum:** the term that the user provides in the query, for which they are seeking a definition;
- **Legislation filter:** can take one of two values: "it" or "eu", which indicate that the user requested a retrieved or generated definition from the Italian or European jurisdiction, respectively. Additionally, the system can operate with no filter, in which case it will prioritize EU legislative definitions.
- **Date filter:** specifies temporal constraints applied by the user to the retrieval of definitions. This component is crucial for a system that operates in the legal domain, where norms are continuously amended and users might be interested in definitions as they existed in a specific point in time. Both open and close intervals can be used (e.g., "2021-10-23 - 2021-12-31"). Like the legislation filter, this parameter is optional.

The system then retrieves relevant definitions based on embeddings, specifically those produced by the BGE-M3 multilingual model [17]. This model enables the implementation of a hybrid retrieval approach, with a similarity function that incorporates a lexical component and a semantic one. In particular, the semantic (dense) similarity is obtained from the normalized hidden states of their encoder model when processing the "CLS" token. Given two embeddings $e_q$ and $e_r$, then, the similarity is computed as their dot product (cosine similarity):

$$S_d = e_q \cdot e_r \tag{1}$$

BGE-M3 also produces sparse embeddings, using a ReLU activation function applied to the linear projection of the hidden state in correspondence of each word. Then, the similarity between two sentences is obtained from:

$$S_s = \sum_{t \in q \cap p} (w_{qt} \times w_{pt}) \tag{2}$$

Where $p$ and $q$ are two sentences, and $w_{qt}, q_{pt}$ are the sparse embeddings of two words that belong to both sentences. Finally, we obtain the similarity between the user query and each definition as follows:

$$S = w_1 S_d + w_2 S_s \tag{3}$$

Where $w_1, w_2$ were set to 1.0 and 0.7 respectively, to prioritize semantic similarity over lexical similarity, as this is crucial to retrieve semantically similar definitions when the generation of a new definition is required. The multilingual nature of BGE-M3 allowed us to store the Italian and English vectors in the same vector database. From the previous step, we can extract the top 10 legislative definitions in terms of similarity to the user query. Our choice of 10 was motivated by a trade off between providing the model with relevant definitions and to avoid irrelevant ones, and we found 10 to be a good balance in

this context. Once the 10 definitions have been extracted, we filter out those that do not match the specified jurisdiction if required, meaning that we use definitions from EurLex for the "eu" filter, while we use Normattiva and PDL for the "it" one.

The next step in our pipeline is to build a timeline for each definition, containing each version and the range of dates for which it was in force. This is achieved through a query to the eXist XML database, which allows us to retrieve the successive consolidated versions of each document and all the definitions contained within them. Through the timeline of definitions, the system can then apply the temporal filter to the timelines and select only the temporally relevant definitions. The complete timeline is presented to the user, so they can assess which version of a given definition they should adopt. In order to provide the model with a broader context about definitions, we also incorporate metadata of EuroVoc terms associated with the original documents of each retrieved definition in the EurLex corpus as metadata in the Akoma Ntoso documents and stored along with the definitions inside the vector store. The next step in the pipeline is the choice, performed by the model, of which of the retrieved definitions present to the user. We prompt the LLM to adopt the user persona of a legal expert specialized in legislative drafting, with the goal to identify the definitions that should be provided to the user from a given list. In cases of ambiguity, the model is instructed to leverage the information from the EuroVoc terms extracted in the previous step. Furthermore, we instruct the model to prefer EU definitions in an attempt to favor the harmonisation of the EU law.

If the LLM does not find any definition to be relevant to the user query, it opts to generate a new one. In the generation prompt, the LLM is asked to adopt the user persona of a legal expert of the relevant jurisdiction, and the 10 most relevant definitions are provided as examples to leverage the model's in-context learning capabilities and mimic in terms of style, tone, structure and lexicon of existing definitions.

Our prompt for the generation of definitions was first drafted with the help of researchers from the legal domain. A small-scale qualitative evaluation of the generated EU legislative definitions was then performed using the procedure outlined in Section 8. This evaluation highlighted a series of issues with the results:

- **Technological neutrality:** the system should not produce definitions that mention specific companies or technologies;
- **Terminology**: in the EU, some terms have very specific meanings. For example, "Country" means a third party country, while "Member State" indicates one of the states that are members of the EU;
- **Invalid References:** if the system generates references, they should be valid and point to existing and in force legislative documents.
- **Usage of examples:** the generated definitions should be as general as possible, without relying on specific examples as the individual instances are subject to interpretation from the judges.

Thanks to this previous evaluation, we were able to modify the generation prompt to give precise instructions that could counteract some of the problematic tendencies of the model when generating definitions. The results of this effort are discussed in Sec. 8.

## 6. Experimental Settings

We evaluate the system using "Llama 3.3 70B Instruct" as the agent model, a relatively recent model from the LLama 3 family [18]. Crucially, it is an open weights and multilin-

gual model which is able to effectively operate as an agent and to provide instructions on how to use tools. The evaluation is carried out using four distinct datasets, each designed to target distinct functionalities of the system through both quantitative and qualitative analyses:

- **LexDrafter Dataset:** originally introduced in the evaluation of the LexDrafter framework [16], contains 1,330 legislative definitions related to 1,007 distinct English terms.
- **Italian Definitions Dataset:** contains 1478 Italian legislative definitions, randomly sampled from the 4,985 entries extracted from the Normattiva and PDL datasets.
- **EU Qualitative Dataset:** this dataset has been curated by legal experts and contains 73 English terms that are not found in any EU legislative documents.
- **IT Qualitative Dataset:** includes Italian equivalents of the 73 terms in the EU Qualitative Dataset, translated manually by legal experts.

### *6.1. Quantitative Generation Evaluation*

A general assessment of the system's generation capabilities is conducted using the "LexDrafter Dataset" and the "Italian Definitions Dataset" to evaluate performances in the generation of English and Italian definitions, respectively. The generated definitions are evaluated using BLEU, ROUGE, BERTScore and BLEURT. For BERTScore, we used "distilbert-base-uncased" model for English definitions and "distilbert-base-multilingual-cased" for Italian ones. For BLEURT, we used the BLEURT-20 checkpoint, which supports both English and Italian. The system is prompted to define each term in the corresponding dataset language using the queries: *"What is the definition of* {*term*} *in the European Legislation?"* for English and *"Qual è la definizione di* {*term*} *nella legislazione italiana?"* for Italian. To prevent the system from retrieving existing definitions from its own corpus, we force their generation by excluding any definitions from the initial retrieval phase (Sec. 5) whose definiendum matches the requested term.

### *6.2. Qualitative Generation Evaluation*

Since quantitative evaluations fail to evaluate crucial aspects such as definitions' legal soundness, internal consistency, and alignment with established legal frameworks, expert-based evaluations are necessary to assess the quality of the generated definitions in the context of the legal domain [19]. The qualitative evaluation of the system was conducted by a team of legal experts, who analyzed the definitions generated for the EU Qualitative Dataset and the IT Qualitative Dataset. Each definition was evaluated according to the following dimensions:

- **Accuracy**: Does the definition describe the correct legal content?
- **Contextual Appropriateness**: Is the definition appropriate for the specific legal domain and context?
- **Completeness**: Does the definition include all the essential elements necessary to fully describe the legal concept?
- **Consistency**: Is the definition internally coherent and free from contradictions?
- **Timeliness**: Does the definition contain up-to-date information?
- **References Accuracy**: Are the references used in the definition valid?

The experts evaluated each generated definition by assigning a score for each di-

mension using a five-point Likert scale, where 1 indicates the lowest level of adherence to the dimension and 5 the highest. In addition to these graded dimensions, two binary criteria were added to the evaluation, in order to detect the presence of critical issues:

- **Hallucination**: Does the definition contain any false or incorrect information, particularly in relation to the EU legislation?
- **Irrelevance**: Does the definition include information that is irrelevant to the specific legal term being defined?

## 7. Quantitative Evaluation

For the comparison between the Italian and European datasets in terms of quantitative metrics, it must be noted that while BLEU, ROUGE and BLEURT can in principle be compared between these two datasets, BertScore cannot, as it uses two different models for the two datasets. The first interesting finding when comparing the Italian and European dataset is the fact that the model seems to produce better scores for Italian than for the EU definitions for the comparable metrics. This might be due to the more formulaic nature of norms in the Italian legal tradition, which might produce longer standard sequences of words, which in turn improve the performance of our model.

In the comparison between our approach and the results obtained from LexDrafter using LLaMA-2 and Vicuna, we obtain the best scores for the BLEU and BLEURT metrics, while in terms of BERTScore our model produces comparable results for the Recall value and slightly lower Precision and F1. These results show that our approach is at the very least comparable with LexDrafter, albeit with a discrepancy when applying various metrics. While a small amount of deviation can happen when applying the same model across different library versions and architectures, the Precision and F1 differences should be outside of the margin of error for both BertScore and BLEURT, meaning that there is a significant difference within these metrics. The most probable explanation is that our definitions, while they seem to be more lexically aligned with the original ones, might be slightly more semantically divergent from the point of view of the distilBERT model, even if the BLEURT scores do not reflect this trend. The ROUGE scores for the LexDrafter models are not included in the original publication, so they were not included in the table.

| **Model** | **BLEU** | | | | **ROUGE** | | | **BERTScore** | | | **BLEURT** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 1 | 2 | L | P | R | F1 | |
| DefAgent-IT | 0.34 | 0.26 | 0.22 | 0.19 | 0.35 | 0.21 | 0.31 | 0.85 | 0.87 | 0.86 | 0.51 |
| DefAgent-EU | **0.26** | **0.16** | **0.12** | **0.09** | **0.35** | **0.19** | **0.31** | 0.78 | **0.81** | 0.79 | **0.49** |
| LD-Vicuna | 0.25 | 0.13 | 0.07 | 0.04 | – | – | – | **0.83** | **0.81** | **0.82** | 0.47 |
| LD-LLama-2 | 0.28 | 0.15 | 0.09 | 0.05 | – | – | – | **0.83** | **0.81** | **0.82** | 0.47 |

**Table 1.** Summary of the automated generation metrics scores for our models (DefAgent-IT and DefAgent-EU for Italian and European documents) and for the two models from LexDrafter (LD-Vicuna and LD-LLama-2).

## 8. Qualitative Evaluation

| Metric | IT | EU |
|---|---|---|
| Accuracy | 3.57 | 3.97 |
| Contextual Appropriateness | 4.32 | 4.28 |
| Completeness | 3.94 | 4.01 |
| Consistency | 4.23 | 4.19 |
| Timeliness | 4.06 | 4.41 |
| References accuracy | 2.25 | 2.16 |

**Table 2.** Average scores for each dimension

| Metric | IT | EU |
|---|---|---|
| Hallucination Rate | 0.10 | 0.08 |
| Irrelevance Rate | 0.10 | 0.04 |

**Table 3.** Hallucination and irrelevance rates

Table 2 shows the aggregate qualitative evaluation for each dimension as rated by the legal experts. Overall, the average level of accuracy is higher for the EU documents. This is not a surprise, given the higher number of training document in English in comparison to Italian. The generated definitions are usually appropriate in context and sufficiently complete, with both jurisdictions achieving satisfactory results in the context. They also maintain inner consistency, meaning that they are not contradictory. A significant gap can be observed in timeliness, with Italian definitions being less time-aware and point-in-time. Occasionally, while in scope and accurate, the legal reference provided for the Italian legislation has been repealed. Despite being instructed to evaluate solely in-force pieces of legislation, the system does not retrieve a valid reference.

High-valued definitions tend to be precise, contextually appropriate, and aligned with relevant legal and technological standards (when necessary), or the domain. An ideal definition clearly addresses the core elements of the concept without being limited to a specific sub-domain. For example, the definition of *Autonomous Trading Agent* demonstrates strong timeliness by accurately describing functionality without tying to tie the definition to a specific computational approach (e.g., Machine Learning). Similarly, the definition of *Gender-based Crime* performs well in terms of consistency, as it includes all relevant gender-related characteristics (”gender, gender identity, gender expression or sex characteristics”) and refers to a specific legal act to ensure consistency with the legal framework. Similarly, in the references accuracy metric, well-drafted definitions cite the exact legal provision, such as a specific article of a regulation, thus ensuring legal clarity and traceability. Unfortunately, as the score show, this is not always the case. Most legal references are too broad (e.g., references to the Treaty of Functioning of the European Union), out of scope (e.g., references to other domains), not in force, hallucinated, redundant or recursive (e.g., using the definiens of the referenced law, which does not define the definiendum). Additionally, in terms of contextual appropriateness, certain definitions fail to reflect the practical implications or real-world application of the term, instead relying solely on abstract or overly legalistic formulations.

The results presented in Table 3 indicate an overall strong performance. Both the hallucination and irrelevance rates remain low across the IT and EU datasets. The hallucination rates of 0.10 (IT) and 0.08 (EU) suggest that the system generates mostly factually grounded content, while the irrelevance rates of 0.10 (IT) and particularly 0.04 (EU) highlight a sufficient degree of contextual relevance.

## 9. Discussion and Future Work

In this article we propose an agentic approach for the generation of legislative definitions in the context of European and Italian legislative documents. This approach leverages retrieved normative references from two datasets, then it decides whether to provide the user with a retrieved or generated definition for a given term. Furthermore, we allow the user to filter the results of their query using temporal and jurisdictional constraints. To the best of our knowledge, this is the first attempt to create a multilingual LLM-based approach for the generation of normative references, which is accompanied by a quantitative and qualitative evaluation of the results. In terms of the research questions discussed in the introduction, we can state that:

- **RQ1**: The retrieval of relevant definitions from a given corpora seems to be quite effective, and our agent is able to decide autonomously whether to generate or retrieve documents with a high level of accuracy.
- **RQ2**: Our agent-based approach is at least comparable existing method which is applied to the English language only. In terms of quantitative measures, achieving SOTA performance for the BLEU and BLEURT metrics.
- **RQ3**: Through our qualitative evaluation of the generated definitions, we understood clear limitations in the definition generation, namely with regards to the correct use of external legal references. Minor issues also concern the overall accuracy of the legal content.

In terms of future work, one of the main outcomes of the qualitative evaluation is the problematic nature of generated normative references within definitions. In this context, it would be useful to provide the model with some form of post-hoc evaluation of the references that it produces, allowing the LLM to correct its own output and to avoid the generation of references that are invalid, that point to amended or repeal portions of documents. This would require a careful implementation, as the point-in-time nature of the law means that references that are valid now might point to abrogated parts of documents in the future. This would also require the application of the approach to datasets that contain the entire collection of European and Italian legislative documents, as the validation of references requires access to all relevant legislative documents.

Another avenue might be a fine-tuning procedure that is specifically aimed at the generation of normative definitions. This would allow the model to learn the specific language used in legislative definitions, and avoid some of the issues that emerge with the contextual appropriateness of the results. However, this would not entirely solve the issues that arise from the generation of incorrect normative references.

The type of definitions that are used both for retrieval and for generation in our pipeline should also be expanded, as our model is only leveraging constitutive definitions, while other definitions (e.g., the interpretative ones) could also be included by keeping into account interpretations from the relevant courts in a given jurisdiction.

In conclusion, while the application of an agentic, RAG-based approach for the generation and retrieval of legislative definitions can be considered a successful experiments, the peculiarities and challenges that emerge when dealing with legislative documents is still partially unsolved, and further research is required on this task.

## Acknowledgments

This project is conducted with the support of the European Commission funds within ERC HyperModeLex. Grant agreement ID: 101055185.

## References

[1] Chau BK, Livermore MA. Computational legal studies comes of age. European Journal of Empirical Legal Studies, Virginia Public Law and Legal Theory Research Paper. 2024;(2024-40).

[2] Palmirani M. A smart legal order for the digital era: A hybrid ai and dialogic model. Ragion pratica. 2022;(2):633-55.

[3] Bresciani PF, Palmirani M. Constitutional opportunities and risks of AI in the Law-making process. Federalismi It. 2024;2:1-18.

[4] Palmirani M, Sperberg R, Vergottini G, Vitali F. Akoma Ntoso Version 1.0 Part 1: XML Vocabulary. OASIS Standard; 2018. Available from: http://docs.oasis-open.org/legaldocml/akn-core/v1.0/akn-core-v1.0-part1-vocabulary.html.

[5] Vitali F, Palmirani M, Sperberg R, Parisse V. Akoma Ntoso Version 1.0. Part 2: Specifications. OASIS Standard; 2018. Available from: http://docs.oasis-open.org/legaldocml/akn-core/v1.0/akn-core-v1.0-part2-specs.html.

[6] Katz DM, Hartung D, Gerlach L, Jana A, II MJB. Natural Language Processing in the Legal Domain; 2023. Available from: https://arxiv.org/abs/2302.12039.

[7] Dahl M, Magesh V, Suzgun M, Ho DE. Large legal fictions: Profiling legal hallucinations in large language models. Journal of Legal Analysis. 2024;16(1):64-93.

[8] Greco CM, Tagarelli A. Bringing order into the realm of Transformer-based language models for artificial intelligence and law. Artificial Intelligence and Law. 2023:1-148.

[9] Siino M, Falco M, Croce D, Rosso P. Exploring LLMs Applications in Law: A Literature Review on Current Legal NLP Approaches. IEEE Access. 2025;13:18253-76.

[10] Jayakumar T, Farooqui F, Farooqui L. Large Language Models are legal but they are not: Making the case for a powerful LegalLLM. In: Preoțiuc-Pietro D, Goanta C, Chalkidis I, Barrett L, Spanakis G, Aletras N, editors. Proceedings of the Natural Legal Language Processing Workshop 2023. Singapore: Association for Computational Linguistics; 2023. p. 223-9. Available from: https://aclanthology.org/2023.nllp-1.22/.

[11] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20. Red Hook, NY, USA: Curran Associates Inc.; 2020. .

[12] Improving Access to Justice with Legal Chatbots. Stats. 2020;3(3):356-75. Available from: https://www.mdpi.com/2571-905X/3/3/23.

[13] Mamalis ME, Kalampokis E, Fitsilis F, Theodorakopoulos G, Tarabanis K. A Large Language Model Agent Based Legal Assistant for Governance Applications. In: Janssen M, Crompvoets J, Gil-Garcia JR, Lee H, Lindgren I, Nikiforova A, et al., editors. Electronic Government. Cham: Springer Nature Switzerland; 2024. p. 286-301.

[14] Audrito D, Spada I, Mignone R, Sulis E, Di Caro L. Towards semi-automating European legislative harmonisation analysis: A harmonised glossary for LLM-based legal concept detection. Computer Law & Security Review. 2025;58:106171.

[15] Palmirani M, Vitali F, Longo G, Sante ED, Brega A, D'Arpa A, et al. Legal Drafting Supported by AI: Enhancing LEOS. In: Martino SD, Sansone C, Masciari E, Rossi S, Gravina M, editors. Proceedings of the Ital-IA Intelligenza Artificiale - Thematic Workshops. vol. 3762 of CEUR Workshop Proceedings. Naples, Italy: CEUR; 2024. p. 482-7.

[16] Chouhan A, Gertz M. LexDrafter: Terminology Drafting for Legislative Documents Using Retrieval Augmented Generation. In: Calzolari N, Kan MY, Hoste V, Lenci A, Sakti S, Xue N, editors. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Torino, Italia: ELRA and ICCL; 2024. p. 10448-58. Available from: https://aclanthology.org/2024.lrec-main.913/.

[17] Chen J, Xiao S, Zhang P, Luo K, Lian D, Liu Z. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation; 2024. Available from: https://arxiv.org/abs/2402.03216.

[18] AI@Meta. Llama 3 Model Card; 2024. Available from: https://github.com/meta-llama/

llama3/blob/main/MODEL_CARD.md.

[19] Palmirani M, Sapienza S, Ashley K. A hybrid artificial intelligence methodology for legal analysis. BioLaw Journal-Rivista di BioDiritto. 2024;(3):389-409.

# Short Papers

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251595*

# A Systematic Approach to Assess Languages Proposed for Rules as Code

Sofia ALI [a,1], Giovanni SILENO [a] and Tom VAN ENGERS [a]
[a] *Informatics Institute, University of Amsterdam, the Netherlands*

**Abstract.** This paper presents a framework for evaluating approaches to formalizing legal sources. The framework integrates adoption-related criteria, extending the scope of assessment beyond purely technical dimensions. We build upon three contemporary comparison frameworks for normative specification languages, combining the Technology-Organization-Environment framework with grounded theory and adding criteria from other IT fields. As an illustration, we applied the resulting framework to RegelSpraak (RuleSpeak) from the Dutch Tax Administration.

**Keywords.** Rules as Code, Formal Language Comparison, Technology-Organization-Environment, Grounded theory, Technology adoption, RegelSpraak

## 1. Introduction

The rise of digitalization has led to a global increase in Rules as Code (RaC) initiatives aiming to encode regulations and directives into machine-readable formats [1,2], building on work in normative systems [3], on legal expert systems [4,5], and legal drafting requirements [6]. Various comparative frameworks have been proposed for formal languages and relative tools, focusing on logical languages [7], on tools for policy automation [8], or domain-specific languages (DSLs) [9,10], but a shared evaluation method is lacking so far. Studies often use limited and inconsistent/incomparable experiments [11], and focus primarily on formal and computational aspects while overlooking practical adoption. Against this background, we address the following questions: (RQ1) *which evaluation criteria recur in existing frameworks*; (RQ2) *which set of criteria covers adoption aspects relevant to practice*; and (RQ3) *what insights arise from applying the resulting framework.* We illustrate the approach with an evaluation of RegelSpraak, a controlled natural language (CNL) used by the Dutch Tax Administration (DTA).

*Related works* Formal legal languages are structured to translate legal texts into precise, machine-readable representations while preserving legal meaning and traceability to the source [11,12]. As the number of such legal representation languages has expanded, evaluation frameworks have become necessary. Early work [7] distinguished *epistemological*, *formal*, and *operational* criteria, while other studies [13] emphasized *organizational integration*, *maintainability*, and *strategic alignment*. Three contemporary comparative frameworks exist, each representing a unique evaluative viewpoint and a distinct practi-

[1]Corresponding author: s.ali@uva.nl. This work was partly funded by the Dutch Research Council (NWO) for the HUMAINER AI project (KIVI.2019.006).

cal focus. Chun et al. (2024) propose an eight-point comparative framework that extends prior work with usability, explainability, interoperability, and adoption-oriented criteria, and they apply it to legal DSLs [9]. Kaptijn and Klaver (2024) introduce an assessment framework for RaC with five scope categories and eight assessment aspects that integrate technical and user-oriented aspects [8]. Parvizimosaed et al. (2022) propose a ten-criterion framework for formal contract languages, organizing evaluation dimensions into ontology, language features, and analysis capabilities [10].

*Research Gap* Collectively, these three frameworks illustrate various evaluation perspectives, encompassing semantic and usability issues, compliance, contract and rule modeling, and user-centered design. However, none of these frameworks fully accounts for broader socio-technical factors, such as how well a language fits within an organization, how easily it can be adopted, or how it interacts with external conditions. Besides, various criteria are introduced without the support of more general literature, for instance, on the topic of technology adoption. This gap unveils the possibility of introducing a more integrated approach.

## 2. Method

*Methodology* To answer our research questions, we follow a three-fold approach. In an *analysis* phase, we collect the evaluation criteria proposed in recent comparison studies and organize them using the *Technology-Organization-Environment* (TOE) framework, initially developed in [14] and applied in technological innovation literature. In an *analogy/transfer learning* phase, we incorporate adoption-related criteria from other relevant IT fields, such as cloud computing and enterprise IT. In a *synthetic* phase, we apply *grounded theory* [15] to the identified criteria. The grounded theory process, via *open*, *axial*, and *selective* coding, allowed the identification of similarities, conceptual clusters, and broader evaluative dimensions. Through repeated analysis cycles, we resolved conceptual inconsistencies and semantic overlaps for a more coherent and empirically grounded evaluative structure.

As an illustrative example, we apply the resulting framework on RegelSpraak, a CNL based on RuleSpeak [16]. Its long-term use within the DTA for machine-readable tax rules provides a suitable setting to assess relevance.

*Constructing the evaluative framework* Applying this methodology, we first extracted evaluation criteria from three key contemporary studies on formal legal languages [8,9,10], presented in detail in the related works paragraph. We mapped all these criteria into the TOE framework, classifying each as technological, organizational, or related to the external environment. Table 1 presents this mapping. This clarifies that a formal legal language must be technically sound, adoptable by organizations, and suitable for the wider legal service provision environment. Where criteria could belong to multiple dimensions, e.g. *traceability* and *explainability*, we placed each criterion in the dimension it most directly applies.

As a second step, to ensure the framework is comprehensive, we added criteria from RaC-related studies and the broader literature on technology adoption. These additional factors capture organizations' challenges and drivers when adopting new technologies, complementing language-specific criteria. Table 2 shows the added factors within each TOE dimension.

| TOE Dimension | Evaluation criteria | Sources |
|---|---|---|
| Technology | Analysis; Automated Verification; Capable of expressing a wide variety of contract genres; Capable of expressing a wide variety of legislative and regulative genres; Executable Analysis; Equipped with a formal semantics describing the language in terms of its underlying logics; Legal Concepts; Observables; Programming Paradigm; Time Support; Coding; Contract Reparations | [8,9,10] |
| Organization | Affordability; Explainability, Interpretation explanation; Transparency; Syntactically "low code" and user friendly with documentation and integrated development environment (IDE) support; Maturity; User-centered | [8,9] |
| Environment | Compliance; Compliance Monitoring; Easy to integrate; Oriented toward interoperability; Subcontracting; Traceability; Traceable Identifiability; Application-oriented; Open-source implementation available | [8,9,10] |

**Table 1.** Mapping of evaluation criteria from key sources to TOE dimensions

| Criteria per TOE dimension | Sources |
|---|---|
| ***Technology*** | |
| Relative advantage | [17,18,19,20] |
| Complexity | [17,21,19,20] |
| Compatibility | [20,17,21,18] |
| Security concerns | [17,18,20,19] |
| Expressiveness | [1,2] |
| ***Organization*** | |
| Top management support | [22,20,17,19,23,18] |
| Organization size | [20,18,17] |
| Cost savings | [20,17,19] |
| Readability by both legal and technical experts | [1] |
| ***Environment*** | |
| Political influence | [22,24] |
| Regulatory support | [25,26,18] |
| Public interpretability of encoded rules | [1,2] |
| Interoperability | [2,1] |
| Transparency of encoded rules | [2,1] |

**Table 2.** Additional adoption criteria

As a third step, we merged and simplified all criteria into a final framework applying grounded theory. These categories, along with concise definitions, are shown in Table 3. The result is a comprehensive evaluation framework considering both technical and socio-organizational perspectives.

## 3. Case study: RegelSpraak

We applied the final TOE-grounded framework (Table 3) to RegelSpraak to validate the framework's utility. Table 4 presents the assessment per criterion: ✓ (criterion is supported), ✗ (criterion is not supported), and – (criterion is not applicable to the language's scope or design). In case of partial support, we used both symbols ✓✗.

| Grounded Theory Category | Definition |
|---|---|
| ***Technology*** | |
| Formal semantics and foundations | Ensures the language has a rigorous, unambiguous formal basis |
| Expressiveness | Ability to capture a broad range of rule types and domain logic |
| Traceability | Links between code artifacts and their legal source references |
| Time management | Handling of time-dependent aspects and evolving state |
| Reasoning and logic | Advanced logic features for legal reasoning and inference |
| Security | Ensures data protection and prevents unauthorized access |
| Implementation fit | Refers to how easily a method can be understood and integrated within existing systems |
| ***Organization*** | |
| User-centric design | Accessible tooling, designed for usability by stakeholders |
| Efficiency and operational advantage | Practical benefits in development and execution |
| Explainability | Capacity to explain or justify outcomes of norm execution |
| Resource availability and support | Availability of sufficient resources, and leadership support |
| Maintainability and maturity | Stability, ease of updates, and maintenance processes |
| ***Environment*** | |
| Regulatory alignment | Alignment with external policy and regulatory requirements |
| Industry trends, external stakeholder demand | Alignment with broader trends and stakeholder requirements |
| International collaboration | Facilitation of international standards and cooperation |

**Table 3.** Grounded Theory merged categories

The results demonstrate how RegelSpraak performs against the proposed criteria. It has a formal rule-based structure, suited for legal and regulatory domains, and works effectively for internal use by domain experts. However, it lacks formal and computational semantics, which disables automated reasoning and verification, while execution is only supported through *Agile Law Execution Factory (ALEF)* rather than directly in the language. From an organizational perspective, however, RegelSpraak benefits from resource availability and support, and its efficiency and operational advantage are evident in streamlined rule management at the DTA. The language also shows an implementation fit, integrating with *iKnow Rules* and *ALEF*. Given the language's non-executable scope, only security concerns were marked as not applicable. The environmental dimension, however, shows some challenges. Regulatory alignment is partial and oriented to internal goals. While external interest is growing, wider uptake depends on ALEF becoming open source, and there are no international collaborations or cross-national reuse cases. Overall, RegelSpraak shows strong socio-technical characteristics in its domestic context, but the absence of logic automation and direct execution semantics constrains broader applicability or transferability.

## 4. Conclusion

This paper presents a structured comparison framework for formal legal languages. The framework combines criteria from previous studies, the TOE framework, and related IT literature, and is constructed by applying grounded theory. For RQ1, we observed that contemporary frameworks consider adoption-focused elements, including usability, in-

| Grounded Theory Category | RegelSpraak |
|---|---|
| ***Technology*** | |
| Formal semantics and foundations | ✓✗ Provides a structured and unambiguous syntax, but lacks an explicit semantics as in logic-based languages. |
| Expressiveness | ✓✗ Supports both fiscal and general rule modeling, though not as broadly expressive as formal logic languages. |
| Traceability | ✗ Provides no inherent mechanism for linking code artifacts to legal sources, with traceability achieved through supporting methodology (Wetsanalyse) and tool (iKnow Cognitatie). |
| Time management | ✓✗ Supports specification of time-dependent rules, with execution provided through ALEF. |
| Reasoning and logic | ✗ Allows specification of repair and consistency rules that handle violations, but lacks a full reasoning engine; execution depends on ALEF. |
| Security | – Not applicable, as it does not handle directly execution-level access or data control |
| Implementation fit | ✓ Integrates into IT systems like ALEF and iKnow Rules; low perceived complexity |
| ***Organization*** | |
| User-centric design | ✓ Natural language structure with low-code logic improves usability |
| Efficiency and operational advantage | ✓ Proven reductions in manual processing and improved decision- making |
| Explainability | ✓✗ Produces structured, human-readable rule interpretations for both internal and external users; execution is however supported through ALEF rather than directly in the language. |
| Resource availability and support | ✓ Backed by institutional support and technical expertise |
| Maintainability and maturity | ✓ Long-term operational use; documented updates and institutional anchoring |
| ***Environment*** | |
| Regulatory alignment | ✓✗ Primarily focused on national/internal goals; not aligned with external policy frameworks |
| Industry trends, external stakeholder demand | ✓✗ Increasing interest from external stakeholders and public agencies, though wider adoption depends on ALEF becoming open source. |
| International collaboration | ✗ No current collaborations or cross-national reuse cases |

**Table 4.** Evaluation of RegelSpraak against the TOE framework

teroperability, explainability, and alignment with user needs. For RQ2, we observed that mapping these to TOE provided a balanced view but revealed gaps that necessitated additional criteria from adjacent IT adoption literature, reorganized using grounded theory. For RQ3, applying the refined framework to RegelSpraak revealed socio-technical factors and technical limits that earlier comparisons overlooked. The framework, therefore, offers a practical, wider basis for evaluation. In the future, we will consolidate the operationalization of criteria, applying our framework to multiple formal legal knowledge representation languages, and conduct additional iterations to test and refine its structure.[2]

[2]We acknowledge, for instance, that the specific scope/problem domain targeted by the tools under scrutiny is currently underdeveloped. This issue derives from the TOE framework, for which the scope remains implicit in the "relative advantage" category, always defined with respect to a certain problem domain.

## References


[1] Rapson J, McLaughlin PA, Wong MW, Morris J, Pope R, Schneider M, et al. Rules as Code for a More Transparent and Efficient Global Economy. T7 Task Force 2: Digitalization of the Global Economy, Policy Brief. 2025 April:1-20.

[2] Mohun J, Roberts A. Cracking the code: Rulemaking for humans and machines. OECD Working Papers on Public Governance. 2020;42(1-109).

[3] Sergot MJ, Sadri F, Kowalski RA, Kriwaczek F, Hammond P, Cory HT. The British Nationality Act as a logic program. Communications of the ACM. 1986;29(5):370-86.

[4] Prakken H, Sartor G. Modelling reasoning with precedents in a formal dialogue game. In: Judicial applications of artificial intelligence. Springer; 1998. p. 127-83.

[5] Bench-Capon T, Sartor G. A model of legal reasoning with cases incorporating theories and values. Artificial Intelligence. 2003;150(1-2):97-143.

[6] Coupette C, Hartung D, Beckedorf J, Böther M, Katz DM. Law smells: Defining and detecting problematic patterns in legal drafting. arXiv preprint arXiv:211011984. 2021.

[7] Fensel D, Harmelen FV. A comparison of languages which operationalize and formalize KADS models of expertise. The Knowledge Engineering Review. 1994;9(2):105-46.

[8] Kaptijn B, Klaver J. Assessment of Rules as Code methods concerning the goal of assisting citizens personally with regulations. Digicampus voor GovTech4All; 2024.

[9] Chun A, Wong MW, Lauritsen M. Domain-Specific Languages and Legal Applications. The Journal of Robotics, Artificial Intelligence & Law. 2024;7(1):19-42.

[10] Parvizimosaed A, Sharifi S, Amyot D, Logrippo L, Roveri M, Rasti A, et al. Specification and analysis of legal contracts with Symboleo. Software and Systems Modeling. 2022;21(6):2395-427.

[11] Novotná T, Libal T. An Evaluation of Methodologies for Legal Formalization. In: Explainable and Transparent AI and Multi-Agent Systems. Cham: Springer International Publishing; 2022. p. 189-203.

[12] Bench-Capon TJM, Coenen FP. Isomorphism and legal knowledge based systems. Artificial Intelligence and Law. 1992 Mar;1(1):65-86.

[13] Schuwer RV. Het nut van kennissystemen; 1993.

[14] Tornatzky LG, Fleischer M. The Processes of Technological Innovation. Lexington Books; 1990.

[15] Tie YC, Birks M, Francis K. Grounded theory research: A design framework for novice researchers. SAGE open medicine. 2019;7(1):2050312118822927.

[16] Corsius M, Hoppenbrouwers S, Lokin M, Baars E, Cappellen GSV, Wilmont I. RegelSpraak: a CNL for executable tax rules specification. In: Proceedings of the seventh international workshop on controlled natural language (CNL 2020/21); 2021. .

[17] Al Hadwer A, Tavana M, Gillis D, Rezania D. A systematic review of organizational factors impacting cloud-based technology adoption using technology-organization-environment framework. Internet of Things. 2021;15:10040.

[18] Chittipaka V, Kumar S, Sivarajah U, Bowden JLH, Baral MM. Blockchain Technology for Supply Chains operating in emerging markets: an empirical examination of technology-organization-environment (TOE) framework. Annals of Operations Research. 2023;327(1):465-92.

[19] Park JH, Kim MK, Paik JH. The factors of technology, organization and environment influencing the adoption and usage of big data in Korean firms. 26th ITS Conference. 2015.

[20] Oliveira T, Thomas M, Espadanal M. Assessing the determinants of cloud computing adoption: An analysis of the manufacturing and services sectors. Information & management. 2014;51(5):497-510.

[21] Li L, Min X, Guo J, Wu F. The influence mechanism analysis on the farmers' intention to adopt Internet of Things based on UTAUT-TOE model. Scientific Reports. 2024;14(1):15016.

[22] Adade D, de Vries WT. An extended TOE framework for local government technology adoption for citizen participation: Insights for city digital twins for collaborative planning. Transforming Government: People, Process and Policy. 2024;19(1):53-73.

[23] Zhang H, Xiao J. Assimilation of social media in local government: an examination of key drivers. The Electronic Library. 2017;35(3):427-44.

[24] Adnan HR, Hidayanto AN, Kurnia S. Citizens' or government's will? Exploration of why indonesia's local governments adopt technologies for open government. Sustainability. 2021;13(20):11197.

[25] Gökalp E, Gökalp MO, Çoban S. Blockchain-based supply chain management: understanding the determinants of adoption in the context of organizations. Information systems management. 2022;39(2).

[26] Jans W, Denters B, Need A, Van Gerven M. Mandatory innovation in a decentralised system: The adoption of an e-government innovation in Dutch municipalities. Acta Politica. 2016;51:36-60.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251596*

# Accountability as a Software Engineering Tool for Multi-Agent Systems

Matteo BALDONI [a,1], Cristina BAROGLIO [a], Roberto MICALIZIO [a] and Stefano TEDESCHI [b]

[a] *Università degli Studi di Torino – Dipartimento di Informatica, Torino, Italy*
[b] *Università della Valle d'Aosta - Université de la Vallée d'Aoste, Aosta, Italy*
ORCiD ID: Matteo BALDONI https://orcid.org/0000-0002-9294-0408, Cristina BAROGLIO https://orcid.org/0000-0002-2070-0616, Roberto MICALIZIO https://orcid.org/0000-0001-9336-0651, Stefano TEDESCHI https://orcid.org/0000-0002-9861-390X

**Abstract.** Accountability is a central notion in both political science and software engineering, where it underpins trust, transparency, and responsibility in complex systems. Multi-agent systems (MAS) provide fertile ground for investigating accountability as a first-class computational construct. While earlier approaches in MAS research relegated accountability to an implicit property, recent extensions model it explicitly. This paper explores accountability in the well-known multi-agent JaCaMo framework as an organizational, normative, and structural mechanism. We argue that formalizing accountability in MAS provides a robust foundation for legal reasoning, compliance verification, and transparent governance of complex distributed systems, ensuring traceable responsibilities and structured feedback flows. We introduce its conceptual foundations, the JaCaMo implementation, and its implications for legal knowledge systems, showing how accountability provides a formal bridge between normative theories and executable systems.

**Keywords.** Accountability, Multi-agent systems, JaCaMo.

## 1. Introduction

The notion of *accountability* is deeply rooted in fields like social sciences, political science, and law, where it amounts to a relationship between two individuals, that helps creating expectations on each other's behavior. There is not a universally accepted definition of accountability [1] but, following [2], a defining characteristic of current understanding of accountability is that each of the parties in an accountability relationship recognizes the authorities that are involved in the relationship: one party to exercise particular powers and the other to hold them to account.

In the last fifteen years many organizations (e.g., the United Nations Development Programme [3]) started to adopt "accountability frameworks", which are organization-wide processes for monitoring, analyzing, and improving performance, enabling managers to address systemic issues and incorporate lessons learned into future activities.

[1] Corresponding Author: Matteo Baldoni, e-mail: matteo.baldoni@unito.it.

Such organizational frameworks are supported by formally documented roles, responsibilities, policies, and instruments that distribute authority and reinforce continuous improvement. Inspired by this use of the concept as well as by studies in ethnomethodology [4], we studied how accountability can be transposed to software engineering [5]. Indeed, complex distributed systems, and in particular multi-agent systems (MAS), increasingly operate as socio-technical infrastructures where robustness, transparency, and trustworthiness are essential. In this context, accountability is not only a normative principle but it can also be a *practical design tool* that structures interactions, clarifies responsibilities, and enables agents to dynamically request and provide accounts of their activities, decisions, and obligations. Explicit accountability enhances system robustness and aligns MAS with legal concepts such as duty, liability, justification, and explanation.

This approach is in a way orthogonal to traditional approaches to MAS design which often assume accountability as an *emergent property* of social protocols that regulate interaction, without treating it as a first-class construct [6,7]. While norms and organizational structures provide important context for regulating behavior, the implicit treatment limits the ability to establish system-wide accounting mechanisms and to ensure that critical information reaches the most competent agents.

In this paper, we sketch an account of accountability in MAS, focusing on: (i) the conceptual foundations of accountability as both a normative and structural mechanism, where the account producing and account tackling activities are spread and may involve many agents according to an organizational specification, (ii) its implementation in the JaCaMo framework [8], and (iii) the legal implications of accountability for designing MAS that can serve as trustworthy socio-technical infrastructures.

## 2. The normative and structural dimensions of accountability

We see accountability as a proactive mechanism that allows agents to *retrieve information dynamically*. Such a piece of information is an *account* that an agent (account giver, *a-giver*) reports to another (account taker, *a-taker*) upon request. This is particularly useful when an agent needs information outside its reach to make decisions.

In accountability, we recognize two dimensions [9,10,11]:

- The *normative* dimension creates mutual expectations on the behavior of the involved agents; it captures the *legitimacy*, for the a-taker, of asking (and the *availability* of the a-giver to provide) an account (the standing of the account taker to demand an account).
- The *structural* dimension concerns the *capability* of the a-giver to produce an account; for being held to account about a process, an agent must exert control over it, and must have proper awareness of the situation it accounts for, possibly by relying on other agents.

In a multi-agent organization, tasks are distributed among the agents depending on their roles. We assume that in order to play a role an agent accepts the involved *responsibilities*, that is, it will execute the tasks that are associated with the role and it will pursue the goals that are associated with the role. Of course, the expectation is that an agent that plays a role has the capabilities that the role requires, i.e., it has plans for pursuing the role goals, and for executing the involved tasks.

By exploiting accountability recursively, agents can provide accounts for tasks that were carried out by others, or that involved the intervention of others. To guarantee the reliability of accounts, the recursive structure of accountability should allow tracing accounts back to the agents that are responsible for the involved tasks. In other terms, recursion closes on some responsibility. This explains why accountability cannot be reduced to the right/obligation involved in the relationship (normative dimension). In [10] we introduced a conceptual model that defines organizational accountability by following this line. In [11,5], in turn, a formal model of accountability is proposed.

## 3. Accountability in JaCaMo

We propose an extension to the JaCaMo framework that explicitly encompasses accountability as a first-class element in the design of multi-agent organizations, as well as in agent programming[2]. Briefly, JaCaMo [8] is a conceptual model and programming platform that integrates agents, environments, and organizations. Jason [12] agents are defined by beliefs (knowledge and state), goals (tasks to achieve), and plans (actions triggered by events). The environment is composed of dynamic CArtAgO [13] artifacts which expose observable properties and operations. Moise [14] organizations include roles and groups (structural dimension), global goals and missions (functional dimension), and permissions and obligations binding roles to missions (normative dimension). At runtime, specifications are translated into normative programs in NOPL [15], which regulate the organization by issuing obligations and guiding agents toward shared goals. JaCaMo agents are required to commit to the missions that are associated with their roles, thus they will be required to commit also to the accountability specifications that involve those roles. However, agents are fully autonomous in deciding whether abiding by the (accountability) norms or not. Sanctions can be associated with violations.

*Specifying accountability.* In our extension, organizational specifications include **Notification Policies** that regulate accountability relationships among agents. A notification policy links a target goal to a condition and defines the accounts to be produced when the condition holds. Conditions can range from goal achievement or delay to any custom organizational state. Each policy contains the specification of one or more types of account the agents agree to possibly ask/send. Each specification is made of a set of predicates with arity that agents must instantiate. Such specification acts like a form to be filled by the a-giver. Each policy also includes goals associated with the report:

- **Requesting goals** allow agents to request information when permitted;
- **Accounting goals** oblige other agents to provide the requested information;
- **Treatment goals** (optional) specify how the provided information should be used;
- **Context goals** (opt.) define the conditions under which accounts can be requested.

**Requesting** and **context goals** differ from standard JaCaMo goals because they grant permission rather than impose obligations. Only once an account is requested, the normative system creates obligations for the accounting goals and, similarly when the

[2]The extension is freely available as part of the Moise platform (one of the three platforms that constitute the JaCaMo framework). The source code, together with a comprehensive set of examples, can be found here: `https://github.com/moise-lang/moise`, under the `exceptions` branch.

account is provided, for treatment goals. This framework enables agents to request runtime information from others without embedding such concerns directly in functional goal decompositions of the multi-agent organization. Interestingly, the a-taker will interpret the received accounts in a way that is not possible to the a-giver, who in general will not be aware of the context where the account is useful.

Of course, the accountability design should be sound, that is, it should allow account production. So, for instance, infinitely looping definitions should be avoided.

*Running example: supervising a production line.* Consider a production line that should not be interrupted. A supervisor agent oversees the process and may need to know how many plates remain on the feed belt. Depending on the number, the supervisor can decide whether to continue production, slow it down, or stop it entirely.

A notification policy can specify this accountability relation. The policy would include an account description with arguments such as the number of available plates, a requesting goal requestRemainingStock assigned to the supervisor, and an accounting goal notifyRemainingStock assigned to the feed belt. Additionally, two treatment goals could be defined: slow down production if the remaining stock is between 10 and 1, and stop production if stock reaches zero. By this policy, the supervisor can at any time request stock information by achieving its requesting goal. The feed belt must then provide the account, instantiating the schema with the current number of plates. The supervisor will then be obliged to act on this information according to the treatment goals. This separation keeps the production logic clear while allowing the supervisor to react flexibly to runtime conditions, e.g., by alerting suppliers and avoiding production from stopping.

*Coding a-giver and a-taker.* The supervisor (a-taker) may request the account during its monitoring cycle by achieving goal requestRemainingStock. Once the feed belt responds with a predicate availablePlates(N), the supervisor can use this information directly within its reasoning cycle. Treatment goals encoded in the notification policy oblige the supervisor to broadcast slowdown or stop commands as conditions require.

```
+!superviseCell : not account(stock,_)
    <- // here goes the supervision logic
       if(/* I need to know the remaining stock */) {
         goalAchieved(requestRemainingStock);
         .wait({+account(stock,[availablePlates(N)])});
         !superviseCell;
       }.

+!superviseCell : account(stock,[availablePlates(N)])
    <- // check if the remaining stock N of available plates is ok

+!slowDownProduction
  <- .broadcast(tell,slowdown).

+!stopProduction
  <- .broadcast(tell,stop).
```

The feed belt agent (a-giver), in turn, is responsible for the accounting goal notifyRemainingStock. Upon activation, it counts the plates in storage and issues giveAccount(stock,[availablePlates(N)]). This makes the requested information available to the supervisor through the organizational infrastructure.

```
+!conveyPlateToTable
    <- // steer the feed belt to move the plate

+!notifyRemainingStock
    <- .countPlatesInStorageRack(N);
       giveAccount(stock,[availablePlates(N)]).
```

*Variants of accountability.* Notification policies are flexible and admit several variants. Designers can, thus, enforce strict organizational treatment of accounts, leave decisions to agents, or introduce conditions that control when accountability applies. For instance, a policy may omit treatment, meaning that the a-taker will decide autonomously how to use it. In the example, the supervisor might still choose to slow or stop production, but such actions would be part of its own reasoning, not mandated by the organization. Another variant involves restrictions on when requesting accounts. For instance, the supervisor may be allowed to ask for stock information only after having obtained authorization from a manager. In this case, getAuthorization goal becomes a precondition.

## 4. Implications for legal knowledge and information systems and conclusions

Explicit accountability in MAS carries significant implications for the formalization of legal knowledge and the operation of information systems, as well. Legal systems often demand auditable records of responsibility. Explicit accountability in MAS ensures that every action can be traced through normative and structural chains to responsible agents. This facilitates automated compliance checks, contractual enforcement, and auditing in digital legal environments (improved *Traceability and compliance*). Accountability mechanisms allow agents to justify their actions on demand, mirroring due process requirements. Unlike opaque black-box systems, accountable MAS are designed to produce structured, verifiable accounts, thereby enhancing trust in automated decision-making (improved *Transparency and explainability*). As discussed in [10,11], accountability contributes to *system robustness*, defined as the invariance of a system under perturbations [16]. In legal contexts, robustness can translate to reliability: even under disruptions, the system can provide defensible justifications, ensuring continuity of legal rights and duties. Related to this, accountability supports the realization of *exception handling mechanisms in distributed systems* [17,18], where the agent detecting an exception may not be the most competent to handle it, and relevant contextual information must be transferred from the exception raiser (a-giver) to the right exception handler (a-taker).

Embedding accountability in MAS opens the way for *accountable digital institutions*. MAS could implement legal contracts, compliance procedures, or regulatory oversight directly within their organizational artifacts. Such systems would not only execute processes, but also continuously monitor and justify them in legally intelligible terms.

Beyond multi-agent programming, explicit accountability modeling has broader significance for the design of distributed socio-technical systems. It supports compliance, monitoring, and auditable governance while preserving organizational flexibility and separation of concerns. In this sense, accountability emerges as a unifying principle that advances both the engineering of resilient distributed intelligent systems and the development of transparent, ethically grounded digital institutions.

## Acknowledgments

The work of Cristina Baroglio, Roberto Micalizio and Stefano Tedeschi is part of the project NODES which has received funding from the MUR –M4C2 1.5 of PNRR with grant agreement no. ECS00000036.

## References


[1] Dubnick M. In: Accountability as a Cultural Keyword. Oxford University Press; 2014. p. 23-38.

[2] Grant RW, Keohane RO. Accountability and Abuses of Power in World Politics. The American Political Science Review. 2005;99(1).

[3] Executive Board of the United Nations Development Programme and of the United Nations Population Fund. The UNDP accountability system, Accountability framework and oversight policy. United Nations; 2008. DP/2008/16/Rev.1.

[4] Garfinkel H. Studies in ethnomethodology. Englewood Cliffs, New Jersey: Prentice-Hall Inc.; 1967.

[5] Baldoni M, Baroglio C, Micalizio R, Tedeschi S. Accountability in multi-agent organizations: from conceptual design to agent programming. Autonomous Agents and Multi-Agent Systems. 2023;37(1):1-37. Available from: https://doi.org/10.1007/s10458-022-09590-6.

[6] Chopra AK, Singh MP. From social machines to social protocols: Software engineering foundations for sociotechnical systems. In: Proc. of the 25th Int. Conf. on WWW; 2016. .

[7] Singh MP. Norms as a basis for governing sociotechnical systems. ACM Transactions on Intelligent Systems and Technology. 2013;5(1):21.

[8] Boissier O, Bordini RH, Hübner J, Ricci A. Multi-agent oriented programming: programming multi-agent systems using JaCaMo. MIT Press; 2020.

[9] Baldoni M, Baroglio C, Micalizio R. Fragility and Robustness in Multiagent Systems. In: Baroglio C, Hubner JF, Winikoff M, editors. Post-Proc. of the 8th International Workshop on Engineering Multi-Agent Systems, EMAS 2020, Revised Selected Papers. No. 12589 in LNAI. Springer; 2020. p. 61-77.

[10] Baldoni M, Baroglio C, Micalizio R, Tedeschi S. Reimagining Robust Distributed Systems through Accountable MAS. IEEE Internet Computing. 2021;25(6). Available from: https://doi.org/10.1109/MIC.2021.3115450.

[11] Baldoni M, Baroglio C, Micalizio R, Tedeschi S. Robustness Based on Accountability in Multiagent Organizations. In: Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems. AAMAS '21. International Foundation for Autonomous Agents and Multiagent Systems; 2021. p. 142-50. Available from: http://www.ifaamas.org/Proceedings/aamas2021/pdfs/p142.pdf.

[12] Bordini RH, Hübner JF, Wooldridge M. Programming Multi-Agent Systems in AgentSpeak Using Jason. John Wiley & Sons; 2007.

[13] Ricci A, Piunti M, Viroli M, Omicini A. In: Environment Programming in CArtAgO. Boston, MA: Springer US; 2009. p. 259-88.

[14] Hübner JF, Sichman JS, Boissier O. Developing Organised Multiagent Systems Using the MOISE+ Model: Programming Issues at the System and Agent Levels. Int J Agent-Oriented Softw Eng. 2007;1(3/4):370-95.

[15] Hübner JF, Boissier O, Bordini RH. A Normative Organisation Programming Language for Organisation Management Infrastructures. In: Coordination, Organizations, Institutions and Norms in Agent Systems V. vol. 6069 of Lecture Notes in Computer Science. Springer; 2009. p. 114-29.

[16] ISO/IEC/IEEE. Systems and software engineering - Vocabulary. 24765:2010(E) - ISO/IEC/IEEE International Standard. 2010.

[17] Baldoni M, Baroglio C, Micalizio R, Tedeschi S. Exception Handling as a Social Concern. IEEE Internet Computing. 2022;26(6):33-40. Available from: https://doi.org/10.1109/MIC.2022.3216272.

[18] Baldoni M, Baroglio C, Micalizio R, Tedeschi S. Exception handling in multi-agent oriented programming. Knowledge Engineering Review. 2025;40. Available from: https://doi.org/10.1017/s0269888925000050.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251597*

# Segmentation and Processing of German Court Decisions from Open Legal Data

Harshil DARJI [a,b,1], Martin HECKELMANN [a], Christina KRATSCH [a] and Gerard DE MELO [b]

[a] *Hochschule für Technik und Wirtschaft Berlin, Germany*
[b] *Hasso-Plattner Institute / University of Potsdam, Germany*
ORCiD ID: Harshil Darji https://orcid.org/0000-0002-8055-1376, Martin Heckelmann https://orcid.org/0000-0003-1696-6849, Christina Kratsch https://orcid.org/0000-0003-0565-1112, Gerard de Melo https://orcid.org/0000-0002-2930-2059

**Abstract.** The availability of structured legal data is important for advancing Natural Language Processing (NLP) techniques for the German legal system. One of the most widely used datasets, Open Legal Data, provides a large-scale collection of German court decisions. While the metadata in this raw dataset is consistently structured, the decision texts themselves are inconsistently formatted and often lack clearly marked sections. Reliable separation of these sections is important not only for rhetorical role classification but also for downstream tasks such as retrieval and citation analysis. In this work, we introduce a cleaned and sectioned dataset of **251,038** German court decisions derived from the official Open Legal Data dataset. We systematically separated three important sections in German court decisions, namely *Tenor* (operative part of the decision), *Tatbestand* (facts of the case), and *Entscheidungsgründe* (judicial reasoning), which are often inconsistently represented in the original dataset. To ensure the reliability of our extraction process, we used Cochran's formula with a **95%** confidence level and a **5%** margin of error to draw a statistically representative random sample of **384** cases, and manually verified that all three sections were correctly identified. We also extracted the *Rechtsmittelbelehrung* (appeal notice) as a separate field, since it is a procedural instruction and not part of the decision itself. The resulting corpus is publicly available in the JSONL format, making it an accessible resource for further research on the German legal system.

**Keywords.** Legal NLP, German legal system, Court decisions, Open Legal Data, Dataset

## 1. Introduction

The increasing volume of legal data and growing digitization of court decisions in Germany has created new opportunities to advance NLP in the legal domain. Large-scale initiatives, such as *Open Legal Data* [1] and *GerDaLIR* [2], provide access to thousands of court decisions across different courts and years. However, the usability of such re-

[1] Corresponding Author: Harshil Darji, Harshil.Darji@HTW-Berlin.de.

sources in Legal NLP is affected by inconsistent formatting, missing structural markers, and varying HTML structures [3,4].

The structure of German court decisions follows a consistent legal writing format, where *Tenor* provides the operative part of the judgment, *Tatbestand* summarizes facts and procedural history, and *Entscheidungsgründe* outline the reasoning. For downstream tasks such as legal case retrieval [5], citation analysis, and rhetorical role classification [6], a reliable separation of these sections is a prerequisite [7]. A lack of clear segmentation can directly affect retrieval, where overlapping reasoning with factual history can lead to false matches, and citation analysis, where the significance of a cited statute depends on whether it appears in the reasoning or the operative part. In Retrieval-Augmented Generation (RAG) systems, section-aware chunking can improve interpretability and prevent the model from mixing argumentative and operative content.

Prior research on German court decisions has also highlighted the importance of structured and machine-readable court decisions for retrieval and classification [8,9,10], but existing datasets still lack consistent section boundaries. In this paper, we introduce a dataset of 251,038 German court decisions derived from the official Open Legal Data data dump. Using a rule-based extraction approach with regular expressions, we separated the three decision sections and also identified statutory and case references to make interlinks within the corpus readily available. To ensure reliability, we used Cochran's formula [11] with a 95% confidence level and 5% margin of error, drawing a random sample of 384 cases that we manually verified. In this sample, **97.40%** ($\pm$1.59%) of the cases were correctly processed, confirming the reliability of the extraction procedure. The dataset is publicly available in JSONL format via Hugging Face datasets[2].

## 2. Related Work

The *Open Legal Data platform* introduced by Ostendorff et al. [1] created a large corpus of publicly accessible German court decisions, with metadata including court identifiers and dates. However, while there is a consistent structure for metadata, the decision texts are stored as heterogeneous HTML and lack standardized markup. Several subsequent corpora have since been derived from this source. Glaser et al. [12] created a collection of approximately 100,000 German court rulings to evaluate summarization methods, segmenting texts into units suitable for training summarization models, though without enforcing legal section boundaries. Wrzalik et al. [2] introduced *GerDaLIR*, an information retrieval benchmark linking 123,000 query passages with 131,000 documents based on citations between decisions. Darji et al. [3] explored semantic similarity between German court decisions and statutory provisions, and later published a dataset of 1,944 manually annotated legal references from German court decisions [13]. While these efforts advanced retrieval and similarity modeling in the German legal domain, they did not provide a fully structured corpus that separates *Tenor*, *Tatbestand*, and *Entscheidungsgründe* across all decisions.

In addition to datasets, other studies have highlighted how adding structural clarity to legal texts enables computational applications. Heckelmann [14] examined how legal agreements can be represented as executable smart contracts, thus defining a legal framework for their machine execution. While the focus is on fitting smart contracts into the

[2]harshildarji/openlegaldata. https://huggingface.co/datasets/harshildarji/openlegaldata

existing framework of civil law, the motivation is the same: *structure is a prerequisite for making legal data usable in downstream tasks*.

## 3. Dataset

We created our dataset from the official Open Legal Data dump (as of *2022-10-18*[3]). While this raw dataset is available in JSONL format, the decision text is embedded within a `content` attribute with varying HTML structures. Our extraction pipeline focuses on separating three sections, normalizing court metadata, and extracting statutory and case references.

### *3.1. Extraction Process*

We begin by parsing the HTML string in the `content` attribute with an HTML parser and iterate over visible elements limited to `p`, `h1`–`h4`, `td`, and a custom `rd` tag. The text is then normalized via whitespace collapsing, and empty and duplicate lines are skipped.

Next, we normalize the court metadata by resolving city and state identifiers using the public APIs from Open Legal Data[4] (endpoints `/api/states/` and `/api/cities/`). Missing entries were set to *Unspecified*.

Then, we perform section boundary detection by identifying specific, fixed-vocabulary headers. These section headers are recognized by two exact, line-level patterns per section, which are applied with case insensitivity and require a full-line match. The patterns are:

- Compact form: `^\s*<marker>\s*:*$`
- Spaced-letter form: `^\s*<m a r k e r>\s*:*$`

The section vocabulary is fixed to `tenor`, `tatbestand`, `entscheidungsgründe`, and `gründe`[5]. For each, we then test both the compact and space-letter variants (for example, `tenor` vs. `t e n o r`). The active section defaults to `tenor` until the first header is encountered and non-header lines are appended to the currently active section. This assumption is consistent with German legal drafting practice, where all court decisions typically begin with *Tenor*.

German court decisions generally follow two drafting patterns. ***Urteile*** (*decisions with a hearing*) provide all three sections explicitly: *Tenor*, *Tatbestand*, and *Entscheidungsgründe*. ***Beschlüsse*** (*decisions without a hearing*), although beginning with *Tenor*, usually contain a *Gründe* section subdivided by Roman numerals, where, *Gründe I* corresponds to the *Tatbestand* and *Gründe II* to the *Entscheidungsgründe*. When no subdivision is present, the entire *Gründe* section is treated as *Entscheidungsgründe*. This logic is consistently applied throughout our extraction pipeline.

Additionally, we also identify and extract the *Rechtsmittelbelehrung*, which provides procedural instructions on available appeals. Although it is part of the published decision,

[3] Open Legal Data dump. See https://static.openlegaldata.io/dumps/de/2022-10-18/

[4] Open Legal Data API. https://de.openlegaldata.io/api/

[5] Included in the vocabulary as required for boundary detection, but later divided into *Tatbestand* or *Entscheidungsgründe*

it is not considered a substantive section of judicial reasoning; therefore, we store it separately in the schema.

Following section segmentation, all collected lines are processed with the Legal Reference Extraction tool[6], which identifies legal citations and categorizes them by type (`law`, `case`). Each court decision is then stored as a single JSON object containing normalized metadata, sectioned text, and extracted references, allowing for direct use without additional preprocessing. A detailed JSON example entry showing this structure is available online[7].

### 3.2. Verification Process

Automatic section segmentation of decision texts can introduce subtle errors, such as misplaced boundaries or partial overlap between sections. To evaluate the reliability of our extraction process, we estimated the necessary sample size for manual verification using Cochran's formula with finite population correction. This approach ensures a statistically representative sample size for categorical evaluations, where each decision is either correctly or incorrectly segmented.

The initial sample size for an infinite population is given by:

$$n_0 = \frac{Z^2 \cdot p \cdot (1-p)}{e^2} = \frac{1.96^2 \cdot 0.5 \cdot 0.5}{0.05^2} \approx 384.16, \tag{1}$$

*where* $Z$ is the critical value for the chosen confidence level ($Z$=1.96 for 95%), $p$ is the estimated proportion of correct extractions (set to 0.5 to maximize variance and yield the most conservative—i.e., largest—sample size estimate), and $e$ is the margin of error (0.05 for 5%). Applying the finite population correction for $N$=251,038 decisions yields

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}} \approx 383.58. \tag{2}$$

*where* $N$ in this context serves as the population size. In practice, we use the conservative rounded sample size of $\mathbf{n} = \mathbf{384}$.

Thus, 384 cases were selected uniformly at random and manually reviewed. Each sampled case was checked to confirm that the three sections (*Tenor*, *Tatbestand*, and *Entscheidungsgründe*) were correctly identified, with correctness defined strictly as the absence of overlap between sections. The manual review confirmed correct segmentation in 97.40% of cases, with errors mainly due to rare formatting irregularities in the HTML. Based on this result, we computed a 95% confidence interval using the normal approximation with finite population correction, which yields $(0.9581, 0.9899)$ for the full dataset of 251,038 decisions. This confirms that the true proportion of correctly segmented cases lies between 95.8% and 98.9%, indicating a high level of reliability for the extraction process.

[6] https://github.com/openlegaldata/legal-reference-extraction
[7] Example entry: https://huggingface.co/datasets/harshildarji/openlegaldata#example-entry

### 3.3. Section Coverage

Table 1 shows the coverage of three sections within our dataset. The *Tenor* is present in 87.7% of cases, while the *Tatbestand* appears in 65.4%, and the *Entscheidungsgründe* in 95.1%. The difference in coverage of *Tatbestand* and *Entscheidungsgründe* reflects differences in drafting practice between decision types, with *Urteile* usually containing both sections explicitly, while in *Beschlüsse*, the factual background is sometimes merged into the reasoning when no subdivision into *Gründe I* and *Gründe II* is provided. The *Rechtsmittelbelehrung*, which provides procedural instructions on available appeals, appears in 8,335 decisions (3.32% of the corpus).

**Table 1.** Section coverage over 251,038 decisions.

| Section | Count |
|---|---|
| Tenor | 220,273 (87.7%) |
| Tatbestand | 164,222 (65.4%) |
| Entscheidungsgründe | 238,666 (95.1%) |

**Table 2.** Structural composition of decisions.

| Structure | Count |
|---|---|
| All three sections | 144,383 (57.5%) |
| Only Tenor + Ent. | 63,720 (25.4%) |
| Only Tenor | 11,388 ( 4.5%) |

In addition, there are **176** decisions (**0.07%** of the total) in which all sections are absent. These correspond to cases where the original `content` field is blank. Table 2 shows how the different sections are combined within the dataset. A majority, 57.5%, of the decisions contain all three sections, while only 25.4% of decisions contain only the *Tenor* and *Entscheidungsgründe*, and a small number of decisions, 4.5%, contain only the *Tenor* section.

## 4. Conclusion and Future Work

In this paper, we presented a dataset of 251,038 German court decisions derived from the Open Legal Data dataset. Our dataset provides a consistent structure by segmenting decisions into *Tenor*, *Tatbestand*, and *Entscheidungsgründe*, addressing the inconsistent formatting and incomplete markers of the original raw HTML. We also evaluated the extraction pipeline using a statistically representative random sample, using Cochran's formula. The manual verification confirmed the reliable separation in approximately 97% of cases. The dataset is available in JSONL format and includes metadata as well as extracted references, making it directly usable for further research without additional preprocessing.

We are also currently using this dataset to build a RAG system for German legal texts. The segmented structure allows our retrieval pipeline to separately index case summaries, statutory references, and reasoning paragraphs, which are then aligned with user queries. Our future work will focus on improving retrieval quality by adding ranking and reranking strategies. We also plan to evaluate the RAG system across subtasks that further demonstrate the value of segmentation. These include statute retrieval, where reasoning passages must be aligned with cited provisions, reasoning coverage, where factual context and arguments need to be distinguished, and interpretability, where users benefit from seeing only the most relevant section of a decision. Finally, we will extend this corpus with additional court decisions as they become available and fine-tune the handling of legal drafting variations.

## References


[1] Ostendorff M, Blume T, Ostendorff S. Towards an open platform for legal information. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020; 2020. p. 385-8.

[2] Wrzalik M, Krechel D. GerDaLIR: A German dataset for legal information retrieval. In: Proceedings of the natural legal language processing workshop 2021; 2021. p. 123-8.

[3] Darji H, Mitrović J, Granitzer M. Exploring Semantic Similarity Between German Legal Texts and Referred Laws. In: International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management. Springer; 2021. p. 37-50.

[4] Darji H, Mitrović J, Granitzer M. Challenges and considerations in annotating legal data: A comprehensive overview. arXiv preprint arXiv:240717503. 2024.

[5] Ma Y, Wu Y, Ai Q, Liu Y, Shao Y, Zhang M, et al. Incorporating structural information into legal case retrieval. ACM Transactions on Information Systems. 2023;42(2):1-28.

[6] Santosh T, Isaia A, Hong S, Grabmair M. Hiculr: Hierarchical curriculum learning for rhetorical role labeling of legal documents. arXiv preprint arXiv:240918647. 2024.

[7] Bambroo P, Adhikary S, Bhattacharya P, Chakraborty A, Ghosh S, Ghosh K. MARRO: multi-headed attention for rhetorical role labeling in legal documents. Artificial Intelligence and Law. 2025:1-30.

[8] Glaser I, Moser S, Matthes F. Improving Legal Information Retrieval: Metadata Extraction and Segmentation of German Court Rulings. In: KDIR; 2021. p. 282-91.

[9] Urchs S, Mitrović J, Granitzer M. Towards classifying parts of german legal writing styles in german legal judgments. In: 2020 10th International Conference on Advanced Computer Information Technologies (ACIT). IEEE; 2020. p. 451-4.

[10] Urchs S, Mitrovic J, Granitzer M. Design and Implementation of German Legal Decision Corpora. In: ICAART (2); 2021. p. 515-21.

[11] Cochran WG. Sampling techniques. john wiley & sons; 1977.

[12] Glaser I, Moser S, Matthes F. Summarization of German court rulings. In: Proceedings of the Natural Legal Language Processing Workshop 2021; 2021. p. 180-9.

[13] Darji H, Mitrović J, Granitzer M. A dataset of German legal reference annotations. In: Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law; 2023. p. 392-6.

[14] Heckelmann M. Smart Contracts. Neue Juristische Wochenschrift (NJW). 2018:504-9.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251598*

# An Ontology-Driven Graph RAG for Legal Norms: A Hierarchical, Temporal, and Deterministic Approach

Hudson DE MARTIM [1]
*Federal Senate of Brazil*
ORCiD ID: Hudson de Martim https://orcid.org/0000-0003-2751-3490

**Abstract.** Standard Retrieval-Augmented Generation (RAG) systems fail in the legal domain, as they are blind to the hierarchical, diachronic, and causal structure of law, leading to anachronistic and unreliable answers. This paper introduces the Structure-Aware Temporal Graph RAG **(SAT-Graph RAG)**, an ontology-driven framework designed to overcome these limitations by explicitly modeling the formal structure and diachronic nature of legal norms. We ground our knowledge graph in a formal, LRMoo-inspired model that distinguishes abstract legal Works from their versioned Expressions. To validate this architecture, we present a qualitative, trace-based evaluation demonstrating how an agentic framework can leverage our model's deterministic primitives to resolve complex temporal and provenance queries that are intractable for standard RAG. The result is a practical and auditable framework for building trustworthy and explainable legal AI systems.

**Keywords.** Graph RAG, Legal Knowledge Graphs, Temporal Modeling, Provenance Reconstruction, Computational Law, Deterministic Retrieval, LRMoo

## 1. Introduction

Retrieval-Augmented Generation (RAG) systems [1], while promising, fail to address the unique challenges of legal texts. Legal norms are not flat documents; they possess a formal hierarchy, a dense web of cross-references, and a diachronic evolution through amendments and repeals. This interplay between structural hierarchy and temporal dynamism is a primary obstacle for standard AI, as a "structurally and temporally-naïve" system cannot deterministically retrieve the version of any legal provision valid on a specific date, leading to factually incorrect answers unacceptable in a high-stakes domain.

To address this, we introduce the Structure-Aware Temporal Graph RAG (SAT-Graph RAG), an ontology-driven framework grounded in LRMoo [2]. It models the legal norm's multi-layered structure, propagates changes efficiently via aggregation, and reifies legislative events as queryable Action nodes, enabling a multi-aspect retrieval where a norm's causal history and metadata become as searchable as its textual content. This paper presents this conceptual blueprint and a proof-of-concept evaluation, demonstrat-

[1]Corresponding Author: Federal Senate of Brazil; E-mail: hudsonm@senado.leg.br.

ing how the architecture—formalized in the SAT-Graph API [3]—enables deterministic answers to complex temporal queries that are intractable for standard RAG.

We review related work in Section 2, detail our framework in Section 3, present a proof-of-concept evaluation[2] in Section 4 and conclude in Section 5.

## 2. Related Work

Research in RAG [1] and Graph RAG [4] confirms that external knowledge improves performance. Legal benchmarks like LegalBench-RAG highlight retrieval as a key bottleneck [5], as naive approaches ignore the structural information from legislation that is crucial for statutory retrieval and QA [6,7]. While research on Temporal Knowledge Graphs (TKGs) has focused on representation and reasoning [8], few works integrate the hierarchical and cross-referential structure of legal texts with robust mechanisms for provenance and historical versioning [9]. Legislative standards like Akoma-Ntoso [10] aim to capture these complexities, confirming that legal representation demands explicit temporal and hierarchical support [11]. However, a systematic integration of hierarchical versioning with temporally-aware, provenance-based retrieval remains a critical gap. Our work addresses this gap by grounding a Graph RAG architecture in a formal ontology, enabling deterministic, temporally-aware retrieval with auditable provenance, inspired by hybrid approaches that fuse structured and textual data to improve explainability [12].

## 3. The Proposed Framework

### *3.1. Ontological Foundation and Graph Construction*

Our framework models the hierarchical, temporal, and linguistic complexities of legal norms using a graph structure grounded in a formal model inspired by the IFLA LRMoo ontology [2], which elevates the Work/Expression distinction to first-class entities. While traditional FRBR, designed for bibliographic records, can be interpreted more restrictively, LRMoo is explicitly designed for interoperability and permits modeling temporal and linguistic variants as distinct Expressions. Following legislative informatics practice, we model each dated version of a norm as a distinct `Expression`. The full diachronic modeling principles are detailed in a foundational work [13], currently available as a preprint; here, we define four primary node types:

**Norm:** An abstract legal creation (e.g., "1988 Brazilian Constitution"). (LRMoo: `Work`)

**Component:** A hierarchical element within a norm (e.g., a title, chapter, or article). (LRMoo: `Work`)

**Temporal Version (TV/CTV):** A language-agnostic "temporal snapshot" of a Norm or Component, capturing its semantic content. (LRMoo: `Expression`)

**Language Version (LV/CLV):** The concrete textual realization of a TV/CTV in a specific language. (LRMoo: `Expression`)

[2] A formal data model and API contract is provided as an OpenAPI 3.0 specification, publicly available at: https://github.com/hmartim/sat-graph-api

Graph construction begins with structure-aware semantic segmentation. This process identifies the abstract structural entities, creating the backbone of **Norm** and **Component** nodes (Figure 1). However, linking the textual content (the *Text Unit*) directly to a timeless Component node would be flawed, as it fails to capture the law's diachronic nature: the abstract Component is permanent, but its textual expression is not.

Our model resolves this by using versioning layers. For the initial enactment and for each subsequent amendment, we instantiate a Component **Temporal Version (CTV)** representing its semantic content on a specific date, and a corresponding Component **Language Version (CLV)** for its concrete wording. The text itself is exclusively associated with the CLV, ensuring every piece of content is unambiguously tied to both a specific semantic state and a linguistic expression (Figure 1). This design efficiently scales to multilingual corpora by reusing the abstract and temporal nodes, only creating new language-specific CLVs and their associated **Text Units** for each translation.

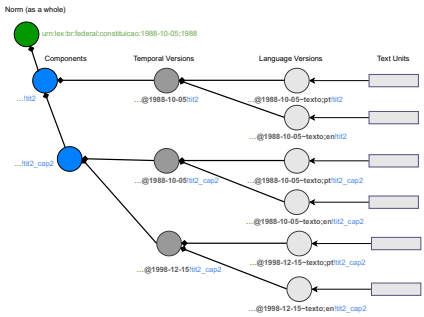



**Figure 1.** The model's layers: Norms (green), Components (blue), Temporal Versions (dark gray), Language Versions (light gray), and Text Units (rectangles).

### *3.2. Temporal Versions as Aggregations*

A key innovation is how our framework propagates changes by replacing naive **Composition** with efficient **Aggregation**. A compositional model, which creates new CTVs for *all* children of an amended component (even those unchanged), is highly inefficient, creates redundant data, and obscures the actual modifications. Our aggregation model for a parent CTV on date $D_n$ simply reuses the *most recent available* CTVs of its unchanged children. This economical approach enables the reuse of child CTVs across parent versions and guarantees the deterministic reconstruction of any provision on a specific date.

### *3.3. Modeling Causality and Metadata as Retrievable Units*

To make the graph's semantic capabilities searchable, we model normative events and metadata as first-class entities. To reify the causality of legislative events, we introduce an **Action** node for each granular change, grounded in event-centric legal ontologies [13, 14]. Each Action represents a specific command from a normative instrument (e.g., an amendment) and acts as a formal causal link, connecting the source provision's original CTV to its effect on the target norm: terminating one CTV/TV and creating another.

We generate a separate, descriptive *Text Unit* for the *Action* itself. This structured summary textualizes the event's key details—such as the amending law, effective dates, and a description of the change. This transforms the legislative event into a searchable object, enabling queries about a law's history and causal evolution, not just its content.

We apply the same principle to structured metadata. Information like publication dates or succession relationships, normally invisible to text-based retrieval, is converted into natural language **Metadata Text Units**, linked to Norm, or Component or Temporal Version, nodes. This operationalizes multi-aspect representation [15], making a norm's entire context—from causal events to metadata—as retrievable as its content.

### *3.4. Structure-Aware Retrieval via Curated Communities*

Departing from standard Graph RAG approaches that rely on algorithmic community detection, our framework leverages two forms of curated communities intrinsic to the legal domain. The first is the norm's own **Internal Hierarchy**: its predefined structure (Titles, Chapters) that provides a natural, nested hierarchy where each grouping component serves as a community. The second is the **External Thematic Classification**: human-curated legal taxonomies (e.g., "Social Security") which are modeled as high-level **Theme** nodes and group related norms and components across the corpus. These nodes can form their own hierarchy, allowing broader themes.

This curated structure enables a powerful filtering mechanism: a user can select any node—a Theme, Norm, Component, or Temporal Version—to define a precise query scope. The semantic search is then performed exclusively on the contextually relevant subset of *Text Units* associated with that scope. This transforms retrieval from a flat search into a process of semantic navigation, yielding more efficient and precise results.

## 4. Proof-of-Concept Evaluation

We validate our framework with a proof-of-concept evaluation using the Brazilian Federal Constitution of 1988, a text with a complex diachronic history. While our conceptual model (Section 3) distinguishes four layers for clarity, this evaluation uses the pragmatic, streamlined model from our SAT-Graph API [3], where Norm and Component are unified into **Item**, the Temporal Version (TV/CTV) is simplified to **Version** and all associated text is collapsed into **TextUnit**. This simplification reduces engineering complexity without compromising the deterministic capabilities demonstrated below.

We present a qualitative, trace-based simulation of an agentic framework resolving complex queries, contrasting its outputs with a baseline representing the dominant paradigm in current RAG deployments: a flat index of a document's current version. Its predictable failures are not presented as a straw man, but serve to illustrate the fundamental architectural limitations our causally-aware framework is designed to overcome.

### *4.1. Modeling a Legislative Amendment: The Case of Article 6*

Our evaluation focuses on Article 6 (more precisely, on its caput), which enumerates social rights and has been amended multiple times. For instance, Constitutional Amendment (CA) No. 26/2000-02-14 added the right to "housing" (*moradia*). In our graph, this is modeled as a discrete Action node that: (1) is causally linked to the Version of the CA

26 provision Item that commanded the change; (2) terminates the original 1988-10-05 Version of Article 6's caput; and (3) creates a new 2000-02-14 Version for Art. 6's caput, linked to a new TextUnit containing the updated pt-BR wording.

### *4.2. Answering Complex Temporal Queries*

*Query Pattern 1: Point-in-Time Retrieval*

For a query like *"What were the social rights listed in Article 6 of the Brazilian Constitution on May 20th, 1999?"*, a standard RAG fails by retrieving the current text, which includes rights added after that date. Our framework succeeds through a deterministic, three-step agent plan: (1) resolve "Article 6, caput of the Brazilian Constitution" to its canonical Item ID (e.g., via the resolveItemReference API call); (2) find the single Version valid on the target date (1999-05-20) (e.g., via getValidVersion); and (3) retrieve its exact TextUnit (e.g., via getTextForVersion). The LLM thus receives only the temporally-correct text, enabling a factually accurate, auditable answer.

*Query Pattern 2: Provenance and Causal-Lineage Reconstruction*

To trace the legislative lineage of a term like "food" in Article 6, a baseline RAG's semantic search is probabilistic and cannot guarantee retrieval of the causal law, nor can it provide a verifiable "before/after" comparison. Our framework enables a deterministic agent plan: (1) resolve "Article 6, caput of the Brazilian Constitution" to its Item ID (e.g., via resolveItemReference); (2) retrieve its full history of Action nodes (e.g., via getItemHistory); (3) identify the pivotal Action that introduced the term by inspecting the text of each resulting Version (e.g., via getBatchTexts). This process yields structured, verifiable data (the amending law, date, and textual difference) for an auditable report.

## 5. Discussion and Conclusion

This paper introduced an ontology-driven Graph RAG architecture that transforms legal retrieval from a probabilistic search into a deterministic, auditable process. Grounded in a formal, LRMoo-inspired model, our principal contributions—explicitly modeling legal structure, versioning, and causal Actions—provide a retrieval substrate superior to flat RAG systems for trustworthy legal AI [16]. While this determinism requires significant data curation, our aggregation strategy demonstrates promising scalability; for a corpus like the Brazilian Constitution, the graph's growth is sublinear relative to the number of amendments—our aggregation strategy for Constitution produced a graph with 5121 Versions (TV/CTVs) across 3085 Items (Norm/Components) modified by 136 Constitutional Amendments. This approach lays a necessary foundation for high-stakes legal AI, though its output quality depends on accurate upstream data.

While this paper's qualitative proof-of-concept validates our deterministic logic, future work will focus on large-scale quantitative evaluation against legal AI benchmarks and the publication of a formal OWL/RDF ontology to support semantic interoperability.

## References


[1] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems. Vol. 33. 2020. p. 9459–74.

[2] IFLA LRMoo Working Group, CIDOC CRM Special Interest Group. LRMoo: Object-oriented definition and mapping from the IFLA Library Reference Model. Version 1.0. IFLA; 2024.

[3] de Martim H. Deterministic Legal Retrieval: An Action API for Querying the SAT-Graph RAG. arXiv preprint arXiv:2510.06002. 2025, doi: 10.48550/arXiv.2510.06002.

[4] Edge D, Trinh HT, Cheng N, Bradley J, Chao A, Mody A, et al. From local to global: A graph RAG approach to query-focused summarization. arXiv preprint arXiv:2404.16130. 2024, doi: 10.48550/arXiv.2404.16130.

[5] Pipitone N, Houir Alami G. LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain. arXiv preprint arXiv:2408.10343. 2024, doi: 10.48550/arXiv.2408.10343.

[6] Louis A, van Dijck G, Spanakis G. Finding the law: Enhancing statutory article retrieval via graph neural networks. arXiv preprint arXiv:2301.12847. 2023, doi: 10.48550/arXiv.2301.12847.

[7] Hei M, et al. A Heterogeneous Graph Based on Legal Documents and Legal Statute Hierarchy for Chinese Legal Case Retrieval. IEEE Access. 2024;12:93502–16.

[8] Cai L, et al. A survey on temporal knowledge graph: Representation learning and applications. arXiv preprint arXiv:2403.04782. 2024, doi: 10.48550/arXiv.2403.04782.

[9] Hooshafza S, et al. Modelling temporal data in knowledge graphs: a systematic review protocol. HRB Open Research. 2022;4:101.

[10] Palmirani M, Vitali F. Legislative XML: principles and technical tools. 2012.

[11] Stellato A, et al. LegalHTML: a Representation Language for Legal Acts. In: ESWC 2023. 2023.

[12] Barron RC, et al. Bridging Legal Knowledge and AI: Retrieval-Augmented Generation with Vector Stores, Knowledge Graphs, and Hierarchical Non-negative Matrix Factorization. arXiv preprint arXiv:2502.20364. 2025, doi: 10.48550/arXiv.2502.20364.

[13] de Martim H. Modeling the Diachronic Evolution of Legal Norms: An LRMoo-Based, Component-Level, Event-Centric Approach to Legal Knowledge Graphs. arXiv preprint arXiv:2506.07853. 2025, doi: 10.48550/arXiv.2506.07853.

[14] de Oliveira Lima JA, Palmirani M, Vitali F. Moving in the time: An ontology for identifying legal resources. In: Legal Knowledge and Information Systems. Berlin: Springer Berlin Heidelberg; 2008.

[15] Park C, Yang C, Zhu Q, Kim D, Yu H, Han J. Unsupervised Differentiable Multi-aspect Network Embedding. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '20). 2020. p. 1435–45.

[16] Richmond KM, et al. Explainable AI and law: An evidential survey. Digital Society. 2024;3(1):1.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251599*

# From Data to Equity: Predicting and Auditing Fairness in Indian Bail Decisions

Sneha DESHMUKH [a,1] and Prathmesh KAMBLE [b]
[a] *Department of Computer Engineering, Datta Meghe College of Engineering, Maharashtra, India*
[b] *Department of Computer Engineering, Datta Meghe College of Engineering, Maharashtra, India*
ORCiD ID: Sneha Deshmukh https://orcid.org/0009-0007-6062-2863, Prathmesh Kamble https://orcid.org/0009-0002-0306-8003

**Abstract.** Algorithmic bias in bail decision systems threatens judicial fairness, yet research on computational analysis of Indian court judgments remains scarce. Building on the recently released *IndianBailJudgments-1200* dataset [4], which includes 1,200 bail cases from Indian High Courts and the Supreme Court (1975–2025) with rich textual and structured socio-legal attributes, we present the first comprehensive empirical study addressing both *bail outcome prediction* and *fairness evaluation* in this context. For bail outcome prediction, we benchmark Logistic Regression and domain-adapted LegalBERT, leveraging combined textual and structured features. For fairness evaluation, we compare Logistic Regression, LegalBERT, and XGBoost across gender, crime type, and region subgroups. Logistic Regression achieves 95.4% baseline accuracy but suffers data leakage (drops to 88.8% after correction); LegalBERT reaches 86.9% with greater robustness. In fairness evaluation, XGBoost attains 95.6% accuracy with lowest gender disparity (1.8%), while LegalBERT achieves 93.9% and Logistic Regression 93.1%. Critically, data leakage correction reveals severe fairness degradation in transformers, with LegalBERT dropping to 82.2%. We will release models, code, and splits, establishing the first reproducible fairness benchmark for Indian bail judgments, and advocate for fairness-aware, human-supervised AI in judicial settings.

**Keywords.** bail outcome prediction, fairness evaluation, LegalBERT, XGBoost, Logistic Regression, Indian judiciary, algorithmic bias, subgroup analysis

## 1. Introduction

Bail decisions profoundly impact accused individuals' liberty, judicial efficiency (court workloads), and public trust in the justice system. In India's overburdened courts, automated analysis enhances consistency and transparency. However, research on Indian bail prediction remains limited due to lack of annotated datasets and language variability.

[1] Corresponding Author: sneha.deshmukh@dmce.ac.in

Recent legal NLP advances show effectiveness globally [1,3,5,2]. Datasets with bail decision fairness evaluation remain rare. IndianBailJudgments-1200 [4] addresses this gap: 1,200 annotated orders from 78 Indian courts (1975–2025) with 20+ structured attributes (crime type, gender, jurisdiction).

**Fairness is critical** [6]: We define *fairness* as **equalized accuracy** across protected subgroups (gender, crime type, region), and *bias* as systematic accuracy disparities. While this operational definition excludes broader judicial fairness aspects, it provides concrete algorithmic equity metrics. **Importantly**, judicial decisions may reflect systemic biases; our labels represent observed patterns requiring scrutiny, not absolute truth.

We investigate: (1) **Bail outcome prediction** using Logistic Regression and LegalBERT; (2) **Fairness evaluation** across all three models (Logistic Regression, LegalBERT, and XGBoost). Results: Logistic Regression 95.4% baseline (drops 88.8% corrected), LegalBERT 86.9% with robustness (85.7% corrected). For fairness: XGBoost 95.6% accuracy, 1.8% gender disparity; LegalBERT 93.9%, Logistic Regression 93.1%. Critically, leakage correction reveals severe fairness degradation—LegalBERT drops to 82.2% (11.7 pp decline).

**Contributions:** (1) First reproducible bail prediction benchmark; (2) Comprehensive fairness audit with rigorous definitions; (3) Critical leakage-correction fairness finding; (4) Open release of trained models, code, splits.

## 2. Dataset

We use IndianBailJudgments-1200 [4]: 1,200 bail orders from 78 Indian High/Supreme Courts (1975–2025), textual summaries, 20+ structured attributes (crime type, region, gender). Corpus: 28 regions, 12 crime categories, 736 Granted/464 Rejected. Notably, 147 landmark, 13 bias-flagged cases support bias detection. Table 1 provides statistics.

**Table 1.** IndianBailJudgments-1200 statistics.

| Attribute | Value |
|---|---|
| Cases | 1,200 |
| Years | 1975–2025 |
| Courts | 78 (High+Supreme) |
| Regions/Crimes | 28 / 12 |
| Outcomes | 736 Granted / 464 Rejected |
| Landmark/Bias-flagged | 147 / 13 |

Reflects typical patterns: predominantly male (class imbalance), 61.3% granted. **Limitations:** Excludes lower courts; female/rare crimes underrepresented. Data split 80/20 train/test (stratified); training split 75/5 validation. Text: lowercasing, tokenization, stopword removal. Features label-encoded. Public archives, anonymized per ethical guidelines.

## 3. Methodology

**Logistic Regression:** TF-IDF (50k vocabulary, unigrams/bigrams), one-hot encoded features; $C \in \{0.1, 1, 10\}$, 5-fold CV, LibLinear solver.

**LegalBERT:** Fine-tuned `nlpaueb/legal-bert-base-uncased`; batch 4, epochs 2, lr $2 \times 10^{-5}$, AdamW, early stopping. [CLS]+16-dim structured embeddings. Hardware (Quadro P500 GPU) necessitated small batches, causing variance ($\pm 14.7\%$), mitigated via seeds (13, 21, 34).

**XGBoost:** Frozen LegalBERT embeddings (768-dim)+structured features; 100 estimators, depth 3, lr 0.1, subsample 0.8, colsample 1.0, $L_2$ $\lambda = 1$, seed 42; 5-fold CV, early stopping.

Data split 80/20 (stratified); 75/5 train/validation. Preprocessing: lowercasing, punctuation/stopword/citation removal (LogReg); minimal cleaning, LegalBERT tokenization (BERT/XGB). Features label-encoded.

**Redaction:** Baseline=full text; Redacted removes "granted," "denied," "rejected," "allowed." **Limitation:** Addresses explicit leakage; metadata/implicit phrasing may persist—future: adversarial validation.

**Fairness:** Max accuracy differences across subgroups (gender, crime, region)=equalized accuracy. Captures algorithmic equity; excludes broader judicial fairness. Differential FP/FN rates may perpetuate injustices.

**Evaluation:** 5 seeds LogReg/XGB (13,21,34,42,57); 3 LegalBERT (constraints). Metrics: accuracy, F1, AUC, Brier, ECE. Paired t-tests ($\alpha = 0.05$), Cohen's $d$, 95% CI. Python 3.10.9; code/data released.

## 4. Results

### *4.1. Bail Outcome Prediction*

Logistic Regression achieves 95.4% baseline but drops to 88.8% after redaction ($t_{(4)} = 23.82$, $p < 0.001$, $d = 10.65$), revealing severe leakage. LegalBERT shows robustness: baseline 86.9% ($\pm 14.7\%$) stabilizes at 85.7% ($\pm 1.8\%$) with AUC improving 8.0 pp, indicating semantic understanding beyond superficial cues [2]. Table 2 summarizes results; Figure 1 illustrates performance trends.

**Table 2.** Bail outcome prediction with statistical tests.

| Model | Condition | Acc. | F1 | AUC |
|---|---|---|---|---|
| LegalBERT | Baseline | $86.9 \pm 14.7$ | $83.1 \pm 19.8$ | $83.9 \pm 21.8$ |
| | Redacted | $85.7 \pm 1.8$ | $84.7 \pm 2.1$ | $91.9 \pm 0.7$ |
| Logistic Reg. | Baseline | $95.4 \pm 1.3^{***}$ | $95.4 \pm 1.3^{***}$ | $99.1 \pm 0.5^{***}$ |
| | Redacted | $88.8 \pm 1.6$ | $88.8 \pm 1.6$ | $96.0 \pm 1.2$ |

*** $p < 0.001$ *for leakage effect. All values in percentages.*

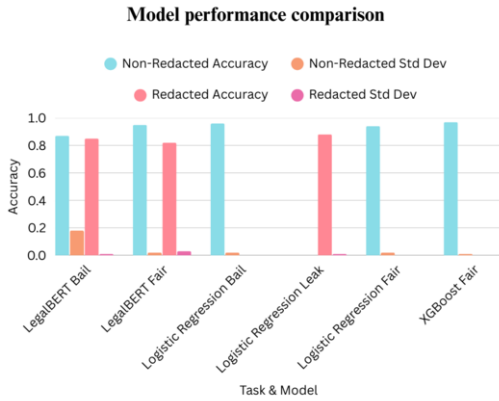



**Figure 1.** Model performance: LegalBERT stabilizes after redaction; Logistic Regression shows leakage-induced degradation.

### *4.2. Fairness Analysis*

XGBoost achieves highest accuracy (95.6%) with lowest gender gap (1.8%). Overall accuracy differences not statistically significant ($F_{(2,10)} = 0.86$, $p = 0.45$), but XGBoost's gender equity shows large practical effect ($d = -1.51$). Crime (15–23%) and regional (24–35%) gaps remain substantial, indicating structural disparities. Table 3 presents subgroup results.

**Table 3.** Fairness: subgroup disparities (max gaps, %).

| Model | Acc. | Gender | Crime | Region |
|---|---|---|---|---|
| LegalBERT (Base) | $93.9 \pm 1.6$ | $2.3 \pm 0.9$ | $22.8 \pm 4.0$ | $34.7 \pm 18.1$ |
| LegalBERT (Red.) | $82.2 \pm 4.1$ | $5.8 \pm 5.2$ | $33.6 \pm 2.9$ | $49.0 \pm 15.7$ |
| Logistic Reg. | $93.1 \pm 0.9$ | $3.9 \pm 3.5$ | $15.0 \pm 2.1$ | $34.7 \pm 17.1$ |
| XGBoost | $95.6 \pm 0.5$ | $1.8 \pm 1.2^{\dagger}$ | $20.0 \pm 5.6$ | $25.0 \pm 7.9$ |

† *Large effect size* ($d > 1.4$).

**Critical Finding:** LegalBERT accuracy drops from 93.9% to 82.2% under redaction, with all gaps worsening (gender: 2.3%→5.8%; crime: 22.7%→33.6%; region: 34.7%→49.0%). This reveals leakage correction can improve generalization while paradoxically compromising fairness—a finding with profound judicial AI implications.

### *4.3. Interpretability*

Logistic Regression coefficients reveal legal terminology ("evidence," "cognizance") and crime severity dominate predictions. XGBoost feature importance shows LegalBERT embeddings contribute 78%, followed by crime type (12%) and region (6%). Classical models rely on explicit keywords (vulnerable to leakage); XGBoost leverages semantic embeddings robustly. Future work should incorporate SHAP/LIME for instance-level interpretability, critical for judicial applications.

## 5. Discussion

LegalBERT's lower baseline (86.9%) but superior robustness (stable 85.7%) indicates deeper semantic understanding versus Logistic Regression's superficial cues. XGBoost combines accuracy (95.6%), fairness (1.8% gender gap), and efficiency (50s training)—offering a practical deployment path.

Crime and regional disparities (15–35%) surpass gender gaps (1–4%), reflecting: (1) differential case complexity; (2) regional legal precedent variations; (3) training underrepresentation. Addressing these requires examining structural judicial inequities, not just algorithmic fixes.

Classical methods show inflated baseline performance (6.6 pp drop); LegalBERT maintains prediction accuracy but suffers severe fairness degradation (11.7 pp). This demonstrates leakage-aware evaluation is essential for safe deployment.

LegalBERT's high variance ($\pm$14.7%) stems from hardware constraints (Quadro P500 GPU) and small batches. This is the main technical limitation. Multiple seed runs partially mitigate instability; training curves show convergence. Future work requires stronger computational resources for stability.

For judicial AI deployment: (1) mandatory leakage-aware evaluation given dramatic fairness impacts; (2) continuous fairness audits of crime/regional gaps; (3) data augmentation for underrepresented groups; (4) human–AI collaboration where AI supports, not replaces, expert judgment. Severe fairness degradation under realistic conditions demonstrates deployment readiness requires evaluation beyond standard accuracy metrics.

## 6. Limitations and Ethics

**Limitations:** (1) Dataset excludes lower courts/unpublished orders, limiting generalizability; (2) Subgroup imbalances reduce reliability for underrepresented groups; (3) Analysis relies on summaries, missing rich contextual data; (4) No temporal dynamics modeled; (5) Limited interpretability—Logistic Regression offers transparency, but LegalBERT/XGBoost require enhanced explainability; (6) No causal inference—statistical associations only.

**Ethics:** Subgroup gaps risk amplifying inequities. Judicial autonomy must remain paramount; AI strictly as decision support. Attribute combinations may enable re-identification despite anonymization. AI recommendations may shift judicial norms; misuse risk exists. Continuous monitoring and stakeholder engagement necessary.

## 7. Conclusion

This study presents a rigorous framework for bail prediction and fairness auditing on IndianBailJudgments-1200. Logistic Regression achieved 95.4% baseline but leakage correction revealed inflated performance; LegalBERT's 86.9% proved more robust. First systematic fairness audit exposed large crime (15–23%) and regional (25–35%) disparities; XGBoost offered best balance (95.6%, 1.8% gender gap). **Critical finding:** Data leakage correction reveals fairness fragility, with transformer performance degrading up to 12 pp—highlighting necessity of robust, leakage-aware evaluation protocols in judicial AI systems. Upon publication, we will release code, models, and data at `https://github.com/[anonymous-for-review]`.

## References

[1] N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, V. Lampos, "Predicting Judicial Decisions of the European Court of Human Rights: a Natural Language Processing Perspective," *PeerJ Comput. Sci.*, vol. 2, e93, 2016, doi: 10.7717/peerj-cs.93.

[2] F. Ariai, J. Mackenzie, G. Demartini, "Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges," arXiv Preprint arXiv:2410.21306, Oct. 2024, doi: 10.48550/arXiv.2410.21306.

[3] I. Chalkidis, I. Androutsopoulos, N. Aletras, "Neural Legal Judgment Prediction in English," in *Proc. 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, Jul. 2019, pp. 4317–4323, doi: 10.18653/v1/P19-1424.

[4] S. Deshmukh, P. Kamble, "IndianBailJudgments-1200: A Multi-Attribute Dataset for Legal NLP on Indian Bail Orders," *arXiv preprint arXiv:2507.02506*, Jul. 2025, doi: 10.48550/arXiv.2507.02506.

[5] S. Paul, A. Mandal, P. Goyal, S. Ghosh, "Pre-trained Language Models for the Legal Domain: A Case Study on Indian Law," in *Proc. 19th Int. Conf. on Artificial Intelligence and Law (ICAIL)*, Braga, Portugal, Jun. 19–23, 2023, doi: 10.48550/arXiv.2209.06049.

[6] T. Y. S. S. Santosh, I. Chowdhury, "Fairness Beyond Performance: Revealing Reliability Disparities Across Groups in Legal NLP," in *Proc. 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Vienna, Austria, Jul. 27–Aug. 1, 2025, pp. 24376–24390, doi: 10.18653/v1/2025.acl-long.1188.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251600*

# Automated Extraction of Judicial Interpretative Formulas in EU Case Law on VAT

Giulia GRUNDLER [a,b,1], Piera SANTIN [c], Alessia FIDELANGELI [a], Rachele MIGNONE [a,d], Federico GALLI [a], Andrea GALASSI [b,1], Giuseppe CONTISSA [a], Luigi DI CARO [d] and Paolo TORRONI [b]

[a] *DSG, University of Bologna*
[b] *DISI, University of Bologna*
[c] *European University Institute*
[d] *Univeristy of Turin*

**Abstract.** This paper addresses the extraction of Judicial Interpretative Formulas (JIFs) in decisions of the Court of Justice of the European Union (CJEU) on Value Added Tax (VAT). European case law includes a significant number of JIFs on this subject, which are crucial for the interpretation of VAT. However, extracting such JIFs manually is effortful, and doing that automatically has not been investigated yet in the VAT domain. Our work proposes the first pipeline method for doing so. We start by defining a set of guidelines for annotating legal texts following a principle definition of JIF. By following such guidelines, we obtain a corpus of 21 expert-labeled CJEU decisions. We keep them for validation and testing. For training, we machine-annotate 80 additional decisions using LLMs. Our experiments show that BERT-based architectures trained on such data perform comparably to LLMs.

**Keywords.** Judicial Interpretative Formulas, Court of Justice of the EU, Value Added Tax, Automated Extraction, Natural Language Processing

## 1. Introduction

In the EU, the Court of Justice of the European Union (CJEU) is responsible for interpreting EU law and ensuring its consistent application. National courts may file requests for preliminary rulings to the CJEU, whose decisions are binding in the referring case. In fulfilling this role, the CJEU adopts an argumentative style with broad interpretative statements whose value extends beyond specific issues, as they are progressively cited and integrated into EU law [1]. Drawing on the literature [2,3], which describes *formulas* as legal texts or standards that the CJEU develops over time through self-citation, we refer to these statements as Judicial Interpretative Formulas (JIFs), highlighting their origin in case law and their interpretative function. These differ from *ratio decidendi* in that they are not limited to the elements necessary to reach the Court's conclusion [4]. One area in which the importance of JIFs is particularly evident is value-added tax. Since VAT

[1] Corresponding Authors: Giulia Grundler: giulia.grundler2@unibo.it, Andrea Galassi: a.galassi@unibo.it.

harmonisation was implemented through a Directive, regulating only certain aspects and leaving others to national discretion, numerous inconsistencies have emerged over time. Consequently, the interpretation of VAT rules has frequently fallen to the CJEU, whose rulings have progressively shaped the substance of VAT law [5].

From a technical perspective, our work builds on the field of automated extraction of legal principles, which, despite their significance, is still a largely manual process, requiring considerable effort from legal practitioners. Recent studies explore the use of automated tools to support this task, relying on hand-crafted grammars [6], machine learning [7,8,9], and prompting LLMs [10]. These approaches face two main drawbacks: data-driven methods require costly annotation, while LLM-based methods often lack stability, transparency, and reproducibility, which are essential in the legal domain.

The objective of this research is to develop and evaluate a machine-learning framework for the extraction of JIFs in CJEU case law on VAT, combining the strengths of data-driven learning with the efficiency of LLMs. We position our solution as a hybrid pipeline that aims to reduce the manual annotation burden and avoid sole reliance on a black-box LLM at deployment. First, we provide a task description that includes defining what constitutes a JIF, resulting in the development of detailed annotation guidelines. Second, we contribute a novel corpus structured into three splits and designed to separate manually annotated data for evaluation and development, and automatically annotated data for training. Third, we present an experimental study evaluating whether state-of-the-art NLP techniques can achieve acceptable results in this field. We observe that BERT models fine-tuned on LLM-annotated data are on par with LLMs in terms of performance, with the potential advantage of more stability, reproducibility, and computational efficiency. The guidelines, corpus, prompts, and code are made publicly available.[2]

## 2. Guidelines, Annotation Process, and Corpus

The source data consists of 101 preliminary rulings of the CJEU in the VAT domain, on the subtopics of taxable amounts and exemptions for the public interest. Of these documents, resulting from a concept-based search using the EUR-Lex Directory[3] in December 2024, 21 were manually labelled, while 80 were labelled automatically.

The annotation guidelines were drafted and iteratively improved following in-depth theoretical and empirical analysis of formulas as used in the argumentative style of the CJEU. As a general rule, we define a JIF as an interpretative statement by the Court, either formulated for the first time or drawn from cited case law. Since the objective is to extract the interpretative statements attributable to the Court, annotation focuses on the part of the decision containing the Court's answer to the preliminary question, which lies between the first occurrence of "Consideration of the question referred" (or similar expressions) and the section on litigant costs (introduced by "Costs"). The annotation unit is one paragraph, reflecting the typical structure of CJEU judgments, which are organised in numbered paragraphs, each expressing an autonomous step of reasoning [11]. In most cases, JIFs have purely legal content, i.e. interpretation of law or legal principles in abstract terms, without any reference to case-specific facts. However, JIFs may occasionally contain factual issues, such as in the final part of the decision, where

[2] https://github.com/poline-project/jif-cjeu/

[3] https://eur-lex.europa.eu/browse/directories/new-case-law.html

**Table 1.** Document sets manually annotated at each stage and corresponding annotation setting.

| Set | Number of documents | Purpose | Annotators per document | Annotators interaction | Split of destination |
|---|---|---|---|---|---|
| A | 2 | guidelines draft | 2 | collaborative | validation |
| B | 9 | guidelines improvement | 2 | independent | validation |
| C | 2 | guidelines validation | 2 | independent | test |
| D | 8 | increase test split | 1 | independent | test |

the Court applies its interpretation of EU law to the case at hand. These passages were also considered JIFs, as they directly express the interpretative stance of the Court. More specifically, we identify as JIFs those paragraphs containing at least one of the following: (i) interpretation of a rule, a part of a rule, or a general principle, (ii) consequences stemming from the interpretation/application of a rule or principle, (iii) subsumption of a fact within a rule, (iv) qualification of a fact as a concept contained within a rule.[4]

### *2.1. Manual Annotation Method*

We created our guidelines and corpus through multiple stages of annotation, each involving different documents and methodologies, as reported in Table 1. The guidelines were envisioned not only as instructions for human annotators, but also as a potential source of information for automated annotation methods. To maximise their effectiveness, the guidelines included examples taken from existing documents. Since this provides a clear advantage for automatic annotation on those same texts, it required careful partitioning of the corpus to ensure sound downstream evaluation. In particular, the test split contains only documents that were never used during the development of the guidelines.

The first draft of the guidelines was based on theoretical and empirical studies. Two annotators jointly worked on a shared set of documents (set A), iteratively refining the guidelines until no ambiguities remained. To validate the guidelines, the annotators then independently annotated a new set of decisions (set B), in a double-blind process. Disagreements were discussed and resolved, leading to a consolidated version of the guidelines. A third set of decisions (set C) was then independently annotated by both annotators and used to validate this final version. We measured the Inter-Annotator's Agreement (IAA) on set C at paragraph level, restricted to the relevant portions of the judgments. The resulting Cohen's $\kappa$ [12] was 0.96, indicating almost perfect agreement.

Since the guidelines were refined after annotating sets A and B, and thus relied on examples from documents contained therein, these sets were excluded from the test split and merged to form the validation split. Conversely, the guidelines remained unchanged after annotating set C, which was included in the test split, further increased by including a fourth set of decisions (set D), with each document annotated by a single expert.

### *2.2. Automated annotation*

Given the large costs associated with manual annotation by legal experts and the need for a large annotated dataset to train BERT-based methods, we decided to automatically annotate additional documents through an LLM. We framed the task as a paragraph clas-

[4] An extended version of the guidelines, including examples, is available in our repository.

**Table 2.** Results of LLMs on the val split.

| Model | Precision | Recall | F1 |
|---|---|---|---|
| Claude | 0.688 | 0.914 | 0.785 |
| DeepSeek | 0.687 | 0.978 | **0.807** |
| Gemini | 0.644 | 0.835 | 0.727 |

**Table 3.** Composition of the dataset.

| Split | N docs | Annot. | JIF | | |
|---|---|---|---|---|---|
| | | | Yes | No | $Yes_{Avg}$ |
| Train | 80 | auto | 1015 | 1037 | 12.7 |
| Val | 11 | manual | 147 | 164 | 13.4 |
| Test | 10 | manual | 142 | 164 | 14.2 |
| Total | 101 | - | 1304 | 1365 | 12.9 |

sification, where the model is prompted to label each paragraph as JIF or non-JIF. To reduce the computational footprint, documents were truncated to the part of the judgment where JIFs occur, as defined in the guidelines. After a preliminary study, we defined a prompt based on our guidelines and evaluated a subset of LLMs: Gemini-1.5-pro, Deepseek-R1, and Claude-3.7-sonnet.[5] Each model was compared against the manually annotated validation split to measure precision, recall, and F1 score (Table 2). DeepSeek obtained the highest score and was therefore adopted to annotate the documents.

A detailed breakdown of the final corpus composition is reported in Table 3.

## 3. Automated extraction

We approached the automatic extraction of JIFs as a binary classification task: given a paragraph, classify it as containing a JIF or not. Experiments were conducted using the train-validation-test splits described in Table 3, determined at the document level so that paragraphs from the same document would never split across partitions. Note that the validation and test splits were entirely manually annotated, ensuring that they maintained the highest quality and avoiding the evaluation of AI models using AI-annotated data.

We experimented with four models: **LinearSVC** with TF-IDF features, **DistilRoBERTa** [13], **DeBERTa** [14], and **LEGAL-BERT** [9]. BERT models were fine-tuned for 10 epochs with early stopping, a learning rate of $2e^{-5}$ and a batch size of 8.[6] With the exception of LinearSVC, which is deterministic, each model was trained three times with different seeds and we report their average performance. For reference, we also evaluated two baselines and **DeepSeek R1**, the generative model used to annotate the training split. The baselines consist of a classifier that outputs a random class (**Random**) and one that always predicts the majority class (**Majority**). For DeepSeek R1, we used the same setting, prompt, and hyperparameters presented in Section 2.2.

Table 4 reports the metrics obtained on the test set by each classifier for each class, as well as their macro-average. The highest macro F1 score of 0.76 was achieved by both DistilRoBERTa and LEGAL-BERT, with all other models closely behind, with a minimum score of 0.72. LEGAL-BERT yielded the best F1 score on the positive class (0.76) followed by DeBERTa and DistilRoBERTa. DeBERTa is the best model for what concerns recall on the positive class, with a score of 0.82, while DistilRoBERTa reaches

[5] We experimented with the following hyperparameters: temperature=0.35, top-k=35, top-p=0.7.

[6] We used the following implementation of the BERT models: *distilbert/distilroberta-base*, *microsoft/deberta-v3-base*, *nlpaueb/legal-bert-base-uncased*.

**Table 4.** Results for the JIF classification task.

| Model | Precision | | | Recall | | | F1 score | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *yes* | *no* | Avg. | *yes* | *no* | Avg. | *yes* | *no* | Avg. | $\sigma$ |
| Majority baseline | 0.00 | 0.54 | 0.27 | 0.00 | 1.00 | 0.50 | 0.00 | 0.70 | 0.35 | - |
| Random baseline | 0.46 | 0.53 | 0.49 | 0.49 | 0.50 | 0.49 | 0.47 | 0.51 | 0.49 | - |
| DeepSeek R1 | 0.73 | 0.78 | 0.75 | 0.73 | 0.76 | 0.75 | 0.74 | 0.77 | 0.75 | - |
| LinearSVC | 0.68 | 0.76 | 0.72 | 0.75 | 0.70 | 0.72 | 0.71 | 0.73 | 0.72 | - |
| LEGAL-BERT | 0.73 | 0.82 | 0.77 | 0.80 | 0.74 | 0.77 | **0.76** | 0.77 | **0.76** | 0.008 |
| DistilRoBERTa | 0.75 | 0.79 | 0.77 | 0.75 | 0.78 | 0.77 | 0.75 | **0.78** | **0.76** | 0.020 |
| DeBERTa | 0.69 | 0.82 | 0.75 | 0.82 | 0.68 | 0.75 | 0.75 | 0.74 | 0.74 | 0.012 |

the maximum precision score of 0.75. For all models except DistilRoBERTa, precision is higher for the negative class while recall is higher for the positive class.

Results indicate that Transformer models generalize well, achieving performance comparable to or even better than DeepSeek. They also exhibit behaviors similar to those observed for the three generative LLMs on the validation set, namely higher recall and lower precision on the positive class, with DeBERTa showing the most pronounced difference. This can be explained considering that the models were trained over data annotated using DeepSeek. Therefore, it is reasonable to speculate that they may have learned similar patterns and a tendency to favour false positives over false negatives. Moreover, the models show stability across repeated runs, as indicated by the low standard deviation of the results. LinearSVC's performance is not much inferior to state-of-the-art models, suggesting that lexical cues play a crucial role in the task.

In general, we consider LEGAL-BERT to be the best model. It is the most stable, and it has the best macro F1 score as well as the best F1 score in the positive class. Moreover, it has the second-best recall score over the positive class, close to the best. This is particularly relevant for tools intended for legal practitioners, as the presence of additional JIFs is preferable to the absence of fundamental ones. Indeed, users can easily discard a few irrelevant paragraphs, but cannot know if a crucial JIF is missing.

## 4. Conclusion

In this paper, we presented a machine-learning framework for the extraction of JIFs from CJEU decisions on VAT cases. We leverage the capabilities of LLMs for annotating training data and explore the performance of various classifiers in this domain.

A general limitation of using LLMs in the legal domain lies in their proprietary nature and lack of transparency, as well as susceptibility to hallucinations [15]. However, our approach mitigates these issues by using a dedicated, task-specific classifier for the final extraction of JIFs, combining the strengths of LLM prompting with a more transparent and reproducible model. Our results indicate that the annotations produced automatically by LLMs can be exploited to distil LLM knowledge in much smaller models that obtain comparable results, reducing annotation costs and improving scalability. By enabling large-scale extraction of these statements, our approach helps various stakeholders, including judges, practitioners, and scholars, in reviewing EU case law.

## Acknowledgments

This work was partially supported by the following projects: POLINE - Principles Of Law In National and European VAT (European Union's Justice Programme, G.A. No.: 101087342); PRIN2022 PRIMA - PRivacy Infringements Machine-Advice (Ref. Prot. n.: 20224TPEYC - CUP J53D23005130001); "FAIR - Future Artificial Intelligence Research" – Spoke 8 "Pervasive AI", under the European Commission's NextGeneration EU programme, PNRR – M4C2 – Investimento 1.3, PE00000013.

## References

[1] Jacob M. Precedents and case-based reasoning in the European Court of Justice: unfinished business. Cambridge University Press; 2014.
[2] Azoulai L. The Retained Powers' Formula in the Case Law of the European Court of Justice: EU Law as Total Law. Eur J Legal Stud. 2011;4:178.
[3] Millet FX. In the name of analogy: Judicial copy-pasting and competence creep in the connection data case law. Common Market Law Review. 2024;61(5).
[4] Komárek J. Reasoning with previous decisions: beyond the doctrine of precedent. The American Journal of Comparative Law. 2013;61(1):149-72.
[5] ARMEANIC A. Added tax: the legal practice of the European Court of Justice securing value. ACROSS Journal of Interdisciplinary Cross-border Studies. 2025 03;7.
[6] Molinari M, Quaranta M, Amantea IA. How do Case Law and Principles of Law interact, Computationally? In: Proceedings of "Workshop on Innovation and Digitization of the Justice System" - Within the XX Conference of the Italian Chapter of AIS - itAIS 2023; 2023. In Press.
[7] Shulayeva O, Siddharthan A, Wyner A. Recognizing cited facts and principles in legal judgements. Artificial Intelligence and Law. 2017;25(1):107-26.
[8] Valvoda J, Ray O. From Case Law to Ratio Decidendi. In: JURISIN. vol. 10838. Springer; 2017. p. 20-34. Available from: https://doi.org/10.1007/978-3-319-93794-6_2.
[9] Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I. LEGAL-BERT: The Muppets straight out of Law School. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics; 2020. p. 2898-904.
[10] Molinari M, Quaranta M, Amantea IA, Governatori G. Using ChatGPT to Extract Principles of Law for the Sake of Prediction: an Exploration conducted on Italian Judgments concerning LGBT(QIA+) Rights. In: Jurix '24 AI for Access to Justice Workshop; 2024. .
[11] Trklja A. The Textual Organisation of CJEU Judgments. International Journal for the Semiotics of Law - Revue internationale de Sémiotique juridique. 2024. Available from: https://doi.org/10.1007/s11196-024-10127-1.
[12] Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement. 1960;20:37 46.
[13] Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv. 2019;abs/1910.01108.
[14] He P, Gao J, Chen W. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing; 2021.
[15] Dahl M, Magesh V, Suzgun M, Ho DE. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. Journal of Legal Analysis. 2024 06;16(1):64-93. Available from: https://doi.org/10.1093/jla/laae003.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251601*

# A Self-Supervised Method for Legal Reference Resolution

Šimon HORVÁT [a], Dávid VARGA [a], Peter GURSKÝ [a], Nicol FEDURCOVÁ [a], Samuel BARAN [a] and Zoltán SZOPLÁK [a]
[a] *Institute of Computer Science, Pavol Jozef Šafárik University in Košice*

**Abstract.** Disambiguating references to legal provisions in court decisions is challenging when abbreviations map to multiple candidate laws, yet it is crucial because reliable reference resolution supports legal information retrieval. We present a self-supervised, graph-based method that constructs a weighted co-occurrence graph from court decisions, where nodes are legal provisions and edge weights reflect their co-occurrence frequency. Ambiguous references are resolved by measuring distances from context provisions using a multi-source Dijkstra algorithm, selecting the closest candidate.

Experiments show that the approach improves accuracy over naive baselines, provides confidence scores for reliability, and delivers a practical disambiguation method, while our study also contributes an annotated dataset for future research.

**Keywords.** legal provision extraction, graph-based disambiguation, co-occurrence graph, court decisions

## 1. Introduction

Parsing references to statutes and legal provisions in legal texts is essential for turning unstructured documents into structured, machine-readable data. In legal texts, references often appear as

| § | ⟨*paragraph*⟩ | ⟨*section*⟩ | ⟨*letter*⟩ | ⟨*act name*⟩ | ⟨*local alias*⟩ |
|---|---|---|---|---|---|
| § | *34* | *section 5* | *letter a)* | *Criminal Code* | *(hereinafter, 'CC').* |

Once an alias is declared, later mentions of the abbreviation (e.g., *CC*) are unambiguous within the document; without an alias, the shorthand may refer to multiple acts—the ambiguity we aim to resolve.

Regular-expression and rule-based methods reliably detect such references and list candidate acts but often fail to disambiguate the abbreviation.

Here, the correct act must be inferred from the surrounding context (e.g., neighboring provisions, recurring terminology, co-occurring references). Existing approaches typically rely on supervised learning with annotated data [1] or on semantic similarity between text embeddings [2].

An alternative is to prompt large language models, but these are resource-intensive. We therefore explore a lightweight method that exploits the structural regularities of legal texts, using LLMs only for low-confidence cases. This hybrid design reduces costs

while maintaining reliable disambiguation. Our self-supervised approach differs by constructing a graph of co-occurring legal provisions and applying shortest-path methods to capture co-occurrence patterns rather than semantic embeddings.

Our contributions are threefold:

- We propose a self-supervised graph-based method, built on co-occurrence patterns of legal provision references, to resolve ambiguous references in court decisions and output prediction confidence scores.
- We evaluate the method on a curated dataset of Slovak court decisions, showing improved accuracy over naive baselines.
- We release the annotated language-agnostic dataset to support further research in legal text processing.

## 2. Related Work

Legal reference extraction has traditionally relied on regular expressions and other rule-based methods; Palmirani et al. (2003) showed that such approaches, which match surface patterns, perform well on well-structured references [3]. Van Opijnen et al. (2015) introduced xLLx, a rule-based system for extracting legal references from Dutch court decisions, which emphasized better maintainability than regular expressions and included handling of "local aliases" of law names [4].

More recent approach introduces machine learning: Darji et al. (2023) frame legal reference prediction as a classification task, using `BERT`-based models to predict the correct law reference from candidate mentions in German court decisions [5].

Graph-based models have also gained traction in legal AI. Zhang et al. (2023) proposed the Missing Link model, a GNN for joint citation prediction combining semantic and topological features [6], and Fujita et al. (2022) explored hypergraphs for hierarchical and self-referential citation modeling [7].

Unsupervised and self-supervised methods are particularly relevant in low-annotation domains. `LamBERTa`, a `BERT` model trained on the Italian Civil Code, leverages unsupervised labeling for extreme classification (i.e., tasks with thousands of possible labels) [8].

While prior graph-based legal work (e.g., Boulet's code networks [9], HierSPCNet [10]) emphasizes document similarity or hierarchy. Our approach adopts a self-supervised paradigm but differs from prior graph-based methods by focusing on co-occurrence graphs of legal provisions, modeling their contextual relationships for disambiguation.

## 3. Dataset Overview

Our experiments are conducted on a dataset of court decisions, specifically designed to support the development and evaluation of legal reference disambiguation. The dataset is divided into two distinct parts: a **graph-building set** (containing only parsed law references without disambiguation) for constructing the co-occurrence graph, and an **evaluation set** (ambiguity-labeled) for testing the disambiguation method. We first describe the evaluation set, so the reader clearly understands the disambiguation task, before detailing how the graph-building set is constructed.

*Evaluation Set* The evaluation set mimics the real-world disambiguation task. Instances were generated by our rule-based extractor [11]: when a local alias was absent and only a law code abbreviation was present, the system consulted a dictionary of possible codes. If multiple candidates existed, the extractor could not resolve the reference. Each instance thus contains the same structural detail (paragraph/section/letter) but multiple candidate codes from this dictionary. Candidates that did not contain the referenced detail in the actual legislation were discarded, ensuring that all remaining alternatives are valid provisions.

The correct provision for each instance was annotated by a domain expert, who inspected the surrounding decision text and selected the provision matching the legal context. This process yielded 123 disambiguation instances.

Each instance represents an ambiguous shorthand reference (e.g., "CC") that could refer to multiple acts containing an identically numbered provision (e.g., both the Criminal and Civil Codes might have a §34). Each includes the correct answer and several plausible, yet incorrect, options. The structure is shown in Table 1.

**Table 1.** Evaluation Set Instance Structure

| Field | Description |
|---|---|
| `laws_before` | A list of legal provisions extracted from the immediate text preceding the ambiguous reference. |
| `laws_after` | A list of legal provisions extracted from the immediate text following the ambiguous reference. |
| `candidates` | A list of all candidate legal provisions for the ambiguous reference. |
| `correct` | The correct, canonicalized legal provision (a single choice from `candidates`). |

The `laws_before` and `laws_after` fields are crucial as they provide the local context for disambiguation. They are sequences of extracted references to legal provisions, ordered exactly as they appear in the court decision text, with the ambiguous reference positioned as a "gap" between them that our algorithm seeks to resolve. The `correct` field, labeled by a legal expert, provides the ground truth for evaluating the accuracy of our method.

**Table 2.** Graph-Building Set Instance Structure

| Field | Description |
|---|---|
| `_id` | A unique identifier for the court decision. |
| `laws` | A list of all unique legal provisions referenced within the decision, transformed to a standard format.<br>Example: `513/1991/paragraph-261.section-3.letter-d` |

*Graph-Building Set* From the evaluation set, we collected all legal provisions that appeared as candidates in any disambiguation instance. For each provision, we retrieved at most 1,000 court decisions containing it; if none were found, the corresponding instance was removed from the evaluation set. This yielded a graph-building corpus of 87,110 court decisions, ensuring that all candidate provisions for disambiguation were repre-

sented in context. This data is used to build the co-occurrence graph, where the frequency of provisions appearing together across court decisions determines the edge weights. The structure of an instance is shown in Table 2.

## 4. Methods

We tackle the legal reference disambiguation problem using a self-supervised, graph-based method that exploits co-occurrence patterns of legal provisions. The approach has two main stages: graph construction and reference disambiguation.

*Graph Construction* The core of our method is a weighted undirected graph $G = (V, E)$ built from the training set:

- **Nodes ($V$)**: Each distinct legal provision occurring in the corpus is represented as a node.
- **Edges ($E$)**: An edge is added between two provisions if they appear together in the same court decision.
- **Edge Weights**: The weight of an edge is the number of co-occurrences of the connected provisions across the corpus. Higher weight therefore indicates a stronger association.

This graph can be constructed efficiently without manual annotation: for each court decision in the graph-building set, we take the set of co-occurring legal provisions from the `laws` attribute and increment the edge weight for every pair.

To improve robustness, we apply edge filtering controlled by a threshold $\tau_{\text{edge}}$: all edges with a co-occurrence count below $\tau_{\text{edge}}$ are removed[1], as they likely reflect noise or rare coincidences.

*Disambiguation* Given the co-occurrence graph, we disambiguate legal references in the evaluation set as follows:

1. **Context Nodes Selection**: From each ambiguous instance, we collect surrounding provisions in the fields `laws_before` and `laws_after`. These context provisions serve as source nodes for distance computation.
2. **Distance Computation**: Using a multi-source variant of Dijkstra's algorithm [12], we compute the shortest-path distance[2] from all context nodes to every other node in the graph.
3. **Score and Select**: For each candidate provision $c$ in `candidates`, we compute the sum of its shortest-path distances from all context nodes:

$$\text{score}(c) = \sum_{\text{ctx} \in \text{context}} \mathit{distance}(\text{ctx}, c).$$

The candidate with the smallest score (i.e., closest to the context) is selected as the predicted resolution.

The scores can be inverted and normalized to yield **confidence** estimates, which quantify the relative closeness of each candidate to the context.

[1] Isolated nodes resulting from this filtering are discarded as well.

[2] Inverse edge weights are used so that paths through stronger connections are favored.

*Matrix Co-occurrence Baseline* As a simpler alternative to our graph-based approach, we implement a matrix-based baseline that relies solely on direct co-occurrence counts. We build a symmetric co-occurrence matrix $M$, where each entry $M_{ij}$ records how often provisions $i$ and $j$ co-occurred in the same decision. During disambiguation, each candidate is assigned a score equal to the sum of its co-occurrence counts with all context provisions, and the candidate with the highest score is selected.

## 5. Experiments and Results

As some instances involve more than two candidates, a random baseline yields 47% accuracy. A classic matrix-based co-occurrence baseline, which scores each candidate by its summed co-occurrence counts, achieves 69% accuracy. Our single predictor based on a co-occurrence graph achieves 76.4%, demonstrating that leveraging indirect connections in the graph improves accuracy. An ensemble of 51 predictors—each built from 500 random decisions—**further improves performance to 79.7%.**

*Error Analysis* Most errors occur when competing provisions frequently co-appear in nearly identical contexts, limiting the signal available from co-occurrence alone. A typical case is the ambiguity between the `513/1991` Commercial Code (lex specialis) and the `40/1964` Civil Code (lex generalis), since commercial law prevails in business matters but defaults to the Civil Code where it is silent.

*Confidence Analysis* As expected, higher confidence correlates strongly with correctness. Predictions with confidence $< 0.7$ achieve 81% accuracy, those with confidence between 0.7 and 0.8 reach 91% accuracy, and predictions with confidence $\geq 0.8$ achieve 97% accuracy. This suggests that confidence can be used to identify highly reliable predictions and flag ambiguous cases for further review.

Implementation and the dataset are available at GitHub repository [13]. All experiments were conducted in Python using Jupyter Notebook, with graph algorithms implemented via the `networkx` library [14].

## 6. Conclusion and Future Work

We presented a self-supervised, graph-based approach for resolving ambiguous legal references in court decisions. By exploiting structural co-occurrence patterns, our method outperforms naive baselines while requiring no manual annotation. Alongside the method, we contribute a curated dataset to support further research.

Future work includes integrating semantic information (e.g., surrounding keywords) with the co-occurrence graph, exploring hybrid scoring strategies, and extending the method to anomaly detection tasks, such as identifying misclassified legal references.

**Acknowledgments:** This research was carried out within the project "Automatization of legal text analysis based on machine learning" funded by the European Union Next GenerationEU through the Recovery and Resilience Plan of Slovakia under project no. 09I05-03-V02-00049.

## References


[1] Ho DE, Grabmair M, Zhang H, Huang Z, Low C, Teng M. Context-Aware Legal Citation Recommendation using Deep Learning. In: Proceedings of the 18th International Conference on Artificial Intelligence and Law (ICAIL); 2021. p. 79-88.

[2] Savelka J, Benyekhlef K, Westermann H. Searching for Legal Documents at Paragraph Level: Automating Label Generation and Use of an Extended Attention Mask for Boosting Neural Models of Semantic Similarity. In: Proceedings of the Natural Legal Language Processing Workshop (NLLP) at ACL-IJCNLP; 2021. .

[3] Palmirani M, Brighi R, Massini M. Automated Extraction of Normative References in Legal Texts. In: Zeleznikow J, Sartor G, editors. Proceedings of the 9th International Conference on Artificial Intelligence and Law, ICAIL 2003, Edinburgh, Scotland, UK, June 24-28, 2003. ACM; 2003. p. 105-6. Available from: https://doi.org/10.1145/1047788.1047815.

[4] van Opijnen M, Verwer N, Meijer J. Beyond the experiment: the eXtendable legal link eXtractor. In: Workshop on Automated Detection, Extraction and Analysis of Semantic Information in Legal Texts, held in conjunction with the 2015 International Conference on Artificial Intelligence and Law (ICAIL); 2015. .

[5] Darji H, Mitrovic J, Granitzer M. A Dataset of German Legal Reference Annotations. In: Grabmair M, Andrade F, Novais P, editors. Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, ICAIL 2023, Braga, Portugal, June 19-23, 2023. ACM; 2023. p. 392-6. Available from: https://doi.org/10.1145/3594536.3595173.

[6] Zhang Z, et al. The Missing Link: Joint Legal Citation Prediction using Heterogeneous Graph Neural Networks. arXiv preprint arXiv:250622165. 2023. Available from: https://arxiv.org/pdf/2506.22165.

[7] Fujita T, et al. A Unified Graph-Based Framework for Modeling Legal Citations: From Graphs to Hypergraphs and SuperHypergraphs. ResearchGate. 2022. Available from: https://www.researchgate.net/publication/394553316_A_Unified_Graph-Based_Framework_for_Modeling_Legal_Citations_From_Graphs_to_HyperGraphs_and_SuperHyperGraphs.

[8] Tagarelli A, Simeri M. LamBERTa: A BERT-based Model for Legal Article Retrieval. arXiv preprint arXiv:211203033. 2021. Available from: https://arxiv.org/abs/2112.03033.

[9] Mazzega P, Bourcier D, Boulet R. The Network of French Legal Codes. Artificial Intelligence and Law. 2011;19(4):333-55.

[10] Bhattacharya P, Ghosh K, Pal A, Ghosh S. Hier-SPCNet: A Legal Statute Hierarchy-based Heterogeneous Network for Computing Legal Case Document Similarity. In: SIGIR '20: Proceedings of the 2020 ACM SIGIR Conference; 2020. Available from: https://arxiv.org/abs/2007.03225.

[11] Varga D, Gojdic M, Szoplák Z, Gurský P, Horvát S, Krajci S, et al. Extraction of Legal References from Court Decisions. In: Brejová B, Ciencialová L, Holena M, Jajcay R, Jajcayová T, Lexa M, et al., editors. Proceedings of the 23rd Conference Information Technologies - Applications and Theory (ITAT 2023), Tatranské Matliare, Slovakia, September 22-26, 2023. vol. 3498 of CEUR Workshop Proceedings. CEUR-WS.org; 2023. p. 89-95. Available from: https://ceur-ws.org/Vol-3498/paper11.pdf.

[12] Xiao Z, Xiao R, Liu C, Gao H, Xu X, Luo S, et al. An Efficient Dynamic Multi-Sources To Single-Destination (DMS-SD) Algorithm In Smart City Navigation Using Adjacent Matrix. arXiv preprint arXiv:221014869. 2022. Available from: https://arxiv.org/abs/2210.14869.

[13] Graph-Based Reference Disambiguation: Implementation and Dataset; 2025. https://github.com/Graph-Based-Reference-Disambiguation/disambiguator.

[14] Hagberg A, Swart P, S Chult D. Exploring network structure, dynamics, and function using NetworkX. Proceedings of the 7th Python in Science Conference (SciPy2008). 2008;836(65-11):11-5.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251602*

# Learning from the Judicial Journey: A Predictive Model for Case Disputability

Ho-Chien HUANG [a✉], Chao-Lin LIU [b]
[a] *Data Science Degree Program, National Taiwan University and Academia Sinica, Taipei, Taiwan*
[b] *Department of Computer Science, National Chengchi University, Taipei, Taiwan*
e-mail: [a]r14946007@ntu.edu.tw, [b]chaolin@g.nccu.edu.tw

**Abstract.** To aid litigants' appeal strategies, this study introduces **Disputability** ($\delta$), a novel metric quantifying a case's contentiousness by the total number of judicial instances it undergoes. Using a dataset of 52,993 Taiwanese tax judgments, we test the hypothesis that similar cases share similar disputability levels. We compare Judgment-Level and Sentence-Level prediction models, finding that a Judgment-Level approach with `multilingual-e5-base` embeddings provides the strongest and most practical baseline. While the Sentence-Level model suffered from significant label noise, a high-precision kNN voting strategy proved particularly effective for identifying high-risk, disputable cases. Our work establishes an adaptable and extensible baseline for quantifying case disputability, offering a practical tool for computational law and legal analytics.

**Keywords.** Case Disputability, Legal Judgment Prediction, Baseline Model, Computational Law, Machine Learning

## 1. Introduction

Whenever a lower court renders a judgment, the losing party invariably faces the decision of whether to appeal. This deliberation is not limited to the loser; the prevailing party, as the potential appellee, must also consider the engagement of legal counsel, which involves an inherent trade-off between litigation costs and potential benefits. Research by [1] suggests that, influenced by psychological effects, litigants are more inclined to retain legal representation only when the probabilities of winning and losing are nearly equal. In other words, the contentious nature, or disputability, of a case impacts a party's consideration of litigation costs, thereby affecting their subsequent litigation behavior.

As for what influences a case's disputability, consider a case that, having been ruled upon by an intermediate appellate court (the second instance), is appealed to the highest court (the third instance), which then reverses and remands the second-instance judgment for further proceedings. This action necessarily implies a divergence in legal opinions between the second and third instance courts; otherwise, the highest court would have affirmed the lower court's decision. Thus, it is evident that a case's disputability has a strong correlation with the number of judicial instances it undergoes. It can even be argued that the disputability of a case is a causal factor influencing the number of instances it passes through.

Based on these observations, we propose a quantifiable measure of case disputability, termed **Disputability**.

**Definition 1.1** (Disputability)**.** Given a case $d_i$, its disputability, denoted as $\delta_i$, is defined as the total number of judicial instances the case has undergone before the final and binding judgment is rendered.

$$\delta_i = \text{Total number of judicial instances for case } d_i$$

For example, if a case undergoes the first instance, second instance, third instance (reverse/remand), second instance (remand proceedings), and is finalized after a subsequent appeal to the third instance (dismissal), it has passed through five judicial instances, and we define its disputability $\delta_i$ as 5.

Considering the practicalities of litigation, the plaintiff and defendant often cite cases with similar facts but differing judicial outcomes as part of their arguments. While the Civil Law tradition may allow for pure dogmatic analysis in the absence of precedent, it is more common to encounter similar cases with somewhat divergent viewpoints. Therefore, the second hypothesis of this study is that similar cases are likely to have similar disputability; in other words, cases similar to a highly disputable case are highly likely to also be highly disputable.

However, the definition of "**similar cases**" has been subject to various discussions. Past research has defined case similarity from multiple perspectives, including legal factors[2][3][4], dialectical reasoning[5][6], and hybrid methods[7], yet no single standard has been universally adopted. This study attempts to utilize three different embedding models without any legal document fine-tuning to define similarity based on Euclidean distance. This approach is intended to serve as a baseline model against which the performance of more specialized legal similarity measures can be evaluated, preserving the exploration of various similarity metrics for this problem. We then examine whether this baseline modeling approach captures similar behavior in terms of disputability, specifically whether highly disputable cases are clustered closer together in the Euclidean space.

## 2. Predictive Model Construction

When analyzing court judgments, regardless of a case's disputability, certain common phrases and boilerplate language are universally present. If judgments are directly embedded, it may be difficult to ascertain which sentences significantly influence the case's disputability. Consequently, we construct two distinct modeling approaches: one based on **Judgment-Level Embedding** for disputability prediction, and the other based on **Sentence-Level Embedding**, followed by aggregation for the final prediction.

Our data source is the Open Data Platform provided by the Judicial Yuan of Taiwan[1]. Due to the practical considerations of applying this research within the Taiwanese legal context[8], we selected tax-related administrative judgments for analysis. The dataset comprises 52,993 judgments and 4,479,889 sentences[2].

[1] `https://opendata.judicial.gov.tw/`

[2] The raw dataset used for this paper are publicly available on the author's GitHub repository: `https://github.com/hochienH/disputability-dataset`

We utilize three different embedding models to capture potential variations: `bert-base-chinese` provided by CKIP Lab[9], `multilingual-e5-base` provided by Microsoft[10], and `bloom-1b1-zh` provided by CKIP Lab and MediaTek[11]. This selection allows us to observe the differences between BERT-based and LLM-based models, and to examine the influence of models pre-trained on Taiwanese Traditional Chinese corpora on the results.

For the judgment-level model, we employ average pooling to transform Sentence Embeddings into Judgment Embeddings[12]. For the sentence-level analysis, each sentence inherits the Disputability ($\delta_i$) of the judgment to which it belongs.

We employ k-Nearest Neighbors (kNN)[13] and Support Vector Machines (SVM)[14] for prediction. For the judgment-level kNN model, we test four neighbor voting strategies (max, mean, median, and majority)[15]. The sentence-level model, in contrast, involves two phases: first, predicting the disputability for each sentence (using the median to mitigate errors), and second, aggregating these sentence-level predictions into a final judgment-level result.

## 3. Results

### *3.1. Judgment-Level Model*

The performance evaluation for the judgment-level prediction models utilizes two fundamental machine learning algorithms: *k*-Nearest Neighbors (kNN) and Support Vector Machines (SVM). All models were trained and validated using **5-fold cross-validation**. Specifically, the kNN model was implemented with the number of neighbors fixed at $k = 5$. The results are evaluated across four key metrics: **Accuracy**, **Precision**, **F1 Score**, and the **F1 Standard Deviation (F1 Std)**.

Table 1 presents the detailed prediction results for the Judgment-Based kNN model ($k = 5$). This table compares the performance across the three chosen embedding models (BERT, Microsoft, and Bloom) under four different voting strategies (Majority, Max, Mean, and Median).

Table 2 summarizes the prediction results for the Judgment-Based SVM model. Unlike kNN, SVM does not require a voting strategy but is similarly evaluated based on the representations provided by the three distinct embedding models.

The comparison across both kNN and SVM algorithms consistently demonstrates the superior performance of the `multilingual-e5-base`. This model achieves the highest overall Accuracy and F1 Scores across the judgment-level prediction tasks. This superior performance may be attributed to its general robustness in common text embedding tasks and the inherent benefit of its multilingual training, which contributes to greater representation power. Consequently, we select the Microsoft model as the primary embedding for the subsequent Sentence-Based experiments to facilitate deeper analysis.

When evaluating the voting strategies within the kNN framework, both the **Majority** and **Median** strategies consistently yield the highest Accuracy and F1 Scores, providing a balanced measure of overall classification performance.

However, an interesting observation is that the **Max strategy**, despite achieving the lowest overall Accuracy, consistently results in the highest Precision. This suggests that

**Table 1.** Performance of the Judgment-Level kNN Model across Embedding Models and Voting Strategies.

| Model | Strategy | Accuracy | Precision | F1 Score | F1 Std |
|---|---|---|---|---|---|
| BERT | Majority | **0.6283** | 0.6115 | 0.6143 | **0.0031** |
| | Max | 0.4719 | **0.7014** | 0.4827 | 0.0032 |
| | Mean | 0.6048 | 0.6221 | 0.6023 | 0.0033 |
| | Median | 0.6255 | 0.6217 | **0.6176** | 0.0033 |
| Microsoft | Majority | **0.6281** | 0.6131 | 0.6159 | 0.0032 |
| | Max | 0.4790 | **0.7083** | 0.4867 | 0.0039 |
| | Mean | 0.6015 | 0.6195 | 0.6010 | 0.0038 |
| | Median | 0.6243 | 0.6219 | **0.6185** | **0.0025** |
| Bloom | Majority | **0.5914** | 0.5737 | **0.5767** | **0.0041** |
| | Max | 0.4026 | **0.6821** | 0.4135 | 0.0050 |
| | Mean | 0.5589 | 0.5862 | 0.5582 | 0.0070 |
| | Median | 0.5834 | 0.5830 | 0.5759 | 0.0048 |

**Table 2.** Performance of the Judgment-Level SVM Model across Embedding Models.

| Model | Accuracy | Precision | F1 Score | F1 Std |
|---|---|---|---|---|
| BERT | 0.6059 | 0.6894 | 0.6326 | **0.0025** |
| Microsoft | **0.6340** | **0.7002** | **0.6512** | 0.0051 |
| Bloom | 0.5813 | 0.6699 | 0.6133 | 0.0067 |

the Max strategy exhibits a high degree of caution when predicting the high-disputability ($\delta_i$) category: once a case is classified as highly disputable, the correctness of that prediction is very high.

In legal practice, litigants may prefer to bear the cost of misclassifying a low-disputability case rather than risk the potentially significant losses from missing a high-disputability case. This makes such a high-precision strategy more practically relevant.

### *3.2. Sentence-Level Model*

Table 3 presents the performance results of the Sentence-Based Prediction Model, which aggregates predictions from the sentence level to the final judgment level using two algorithms: kNN and SVM. We exclusively utilized the best-performing `multilingual-e5-base` embedding model for this experiment. Consistent with the judgment-level analysis, kNN was set with $k = 5$ neighbors, and evaluation was performed using 5-fold cross-validation. The performance metrics evaluated include Accuracy, Precision, F1 Score, and F1 Std. Due to constraints on computational resources, we selected a random sample of 40,000 sentences from the total dataset for testing and validation.

When analyzing the sentence-level predictions, the kNN algorithm demonstrated superior performance compared to SVM. Furthermore, kNN exhibited a more sensitive effect under the **Max strategy**, resulting in a higher Precision score than even the best Judgment-Level models. This highlights the ability of the Sentence-Based kNN model,

**Table 3.** Performance of the Sentnece-Level Model across Algorithms and Voting Strategies.

| Algorithm | Strategy | Accuracy | Precision | F1 Score | F1 Std |
|---|---|---|---|---|---|
| kNN | Majority | 0.6042 | 0.5679 | 0.5159 | 0.0109 |
| | Max | 0.4479 | **0.7324** | 0.3610 | 0.0055 |
| | Mean | 0.5849 | 0.6080 | 0.5710 | 0.0079 |
| | Median | **0.6194** | 0.5709 | **0.5922** | **0.0050** |
| SVM | Majority | **0.5564** | 0.6493 | **0.5371** | 0.0292 |
| | Max | 0.0394 | 0.5972 | 0.0274 | 0.0277 |
| | Mean | 0.4842 | 0.6650 | 0.4482 | 0.0309 |
| | Median | 0.5444 | **0.6665** | 0.5184 | **0.0267** |

when using the Max strategy, to precisely identify highly disputable instances, although at the expense of overall accuracy.

However, its overall performance was significantly lower than the Judgment-Level model's, primarily due to the substantial Label Noise generated by assigning the judgment-level label ($\delta$) to every sentence, which led to an accumulation of errors during aggregation.

## 4. Concluding Remarks

Our study successfully introduced Disputability ($\delta_i$) as a quantifiable metric based on the number of judicial instances a case undergoes. The Judgment-Level models utilizing `multilingual-e5-base` achieved the highest baseline performance, while the Sentence-Based models, despite providing finer granularity, suffered from significant performance degradation due to label noise.

The proposed model is highly extensible: its disputability definition, based on the number of judicial instances, is easily adaptable across different legal domains and systems. The framework can also flexibly incorporate various case similarity metrics to enhance interpretability and performance.

Our baseline model presents several notable limitations. Firstly, the current definition of similarity based on Euclidean distance offers very low legal interpretability and is thus rarely suitable for direct practical use in legal contexts. Secondly, the Sentence-Based model proved overly susceptible to label noise. Effectively leveraging the benefits of this fine-grained processing requires either advanced data sampling techniques or significantly higher computational cost, which currently restricts its practical utility.

Future work should focus on two main areas. Firstly, for the Sentence-Based model, research should be conducted to identify better sampling methodologies that allow the model to perform robustly within smaller sample spaces, while simultaneously exploring various interpretable semantic similarity measures. Secondly, we plan to develop a playground-like interface for the model, providing users with a simpler platform to adjust parameters and construct their own customized quantitative disputability models.

## Acknowledgements

This research was partially supported by the 114-2221-E-004-001-MY2 grant of the National Science Technology Council, Taiwan. The authors sincerely thank the anonymous reviewers for their insightful and constructive feedback, which significantly improved this manuscript. We also appreciate the participation of Chian-Yu Ye, an undergraduate student in the Department of Law, National Chengchi University, Taiwan, in the early discussions of this research.

## References

[1] Zhang HY. Yun-chien Chang, Why Do Litigants in China Choose Not to Hire Lawyers? An Empirical Legal Study Using Chinese Data;. Available from: https://en.law.pku.edu.cn/externalrelations/newsletter/164022.htm.

[2] Ashley KD, Bruninghaus S. A Predictive Role for Intermediate Legal Concepts. In: Bourcier D, editor. Legal Knowledge and Information Systems: JURIX 2003: the Sixteenth Annual Conference. vol. 106 of Frontiers in Artificial Intelligence and Applications. IOS Press; 2003. The full paper is within the 200 page volume, page range for the specific article is needed for 'pages' field.

[3] Aleven V. Using background knowledge in case-based legal reasoning: A computational model and an intelligent learning environment. Artificial Intelligence. 2003 Nov;150(1-2):183-237. Available from: https://linkinghub.elsevier.com/retrieve/pii/S000437020300105X.

[4] Rissland EL, Ashley KD, Branting LK. Case-based reasoning and law. The Knowledge Engineering Review. 2005 Sep;20(3):293-8. Available from: https://www.cambridge.org/core/product/identifier/S0269888906000701/type/journal_article.

[5] Roth B. New reasoning patterns in analogical legal case-based reasoning: an informal investigation. In: Breuker J, Leenes R, Winkels R, editors. Legal Knowledge and Information Systems: JURIX 2000: the Thirteenth Annual Conference. Frontiers in Artificial Intelligence and Applications. IOS Press; 2000. The volume number (likely 69) and page range for this specific article are missing and should be added.

[6] Roth B. The Dialectics of Case Comparison: a Formal Reconstruction. In: Verheij B, Lodder AR, Loui RP, Muntjewerff A, editors. Legal Knowledge and Information Systems: JURIX 2001: The Fourteenth Annual International Conference. vol. 70 of Frontiers in Artificial Intelligence and Applications. IOS Press; 2001. The full paper is within the 150 page volume, page range for the specific article is needed for 'pages' field.

[7] Bhattacharya P, Ghosh K, Pal A, Ghosh S. Legal case document similarity: You need both network and text. Inf Process Manage. 2022;59(6). Available from: https://doi.org/10.1016/j.ipm.2022.103069.

[8] Chen M. Administrative Law: General Principles. 10th ed. New Taipei City: New Xuelin Publishing Co.; 2019.

[9] Yang M, Ma WY, Lab C. ckip-transformers: PyTorch implementation of CKIP models for Chinese NLP tasks; 2023. Version 0.4.0;
Original creation date: 2020-08-17. GitHub repository. Available from: https://github.com/ckiplab/ckip-transformers.

[10] Wang L, Yang N, Huang X, Yang L, Majumder R, Wei F. Multilingual E5 Text Embeddings: A Technical Report. arXiv preprint arXiv:240205672. 2024.

[11] Ennen P, Hsu PC, Hsu CJ, Liu CL, Wu YC, Liao YH, et al.. Extending the Pre-Training of BLOOM for Improved Support of Traditional Chinese: Models, Methods and Results; 2023. Available from: https://arxiv.org/abs/2303.04715.

[12] Goodfellow I, Bengio Y, Courville A. Deep Learning. The MIT Press; 2016.

[13] Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theor. 2006 Sep;13(1):21-7. Publisher: IEEE Press. Available from: https://doi.org/10.1109/TIT.1967.1053964.

[14] Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995 Sep;20(3):273-97. Available from: https://doi.org/10.1007/BF00994018.

[15] Alpaydin E. Introduction to Machine Learning. 4th ed. Adaptive Computation and Machine Learning series. Cambridge, MA: The MIT Press; 2020.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251603*

# Reinforcement Learning with Argument-Structured Reward for Court Decision Abstractive Summarization

Yuntao KONG [a,1], Ye XIONG [b], Shuyuan ZHENG [c] and Ken SATOH [a]
[a] *Center of Juris-Informatics, ROIS-DS, Japan*
[b] *Institute of Science Tokyo, Japan*
[c] *The University of Osaka, Japan*

**Abstract.** Court decision summarization is challenging due to the significant length and structural complexity of legal documents, which makes existing reinforcement learning (RL)-based abstractive summarization methods less effective. We propose an RL-based abstractive legal summarization model with a novel argument-structured reward mechanism. It leverages Issue–Reason–Conclusion components to compute fine-grained sub-rewards and aggregate them into a final reward. This design provides more reliable learning signals for model optimization. Experiments on the Indian Supreme Court dataset with Longformer-Encoder-Decoder (LED) and Llama demonstrate consistent improvements over non-argument-structured methods. To the best of our knowledge, this is the first work to incorporate argumentative structures into RL-based summarization, offering a novel direction for improving legal summarization.

**Keywords.** Court Decision Summarization, Reinforcement Learning, Argument-Structured Reward

## 1. Introduction

Court decision summaries are typically longer and more complex, which challenge current summarization models, particularly for abstractive approaches that aim to generate concise and coherent summaries. Recent advances in reinforcement learning (RL)-based methods have shown promising performance in abstractive summarization. The current reinforcement learning (RL) approach follows a two-stage approach: first, supervised fine-tuning (SFT) is used to obtain a baseline model; then RL training is applied to optimize the model by maximizing pre-defined reward. The reward can directly use evaluation metric scores (e.g., ROUGE, BERTScore) or employ more sophisticated designs [1–3]. However, applying current RL methods to legal summarization reveals significant limitations. Previous work [4, 5] shows that current metric scores are brittle and high-variance for long or structurally complex summaries, thus providing unreliable signals for RL training in this setting. This raises a critical question: How can we design reward

[1] Corresponding Author: Yuntao KONG, kongyt@nii.ac.jp

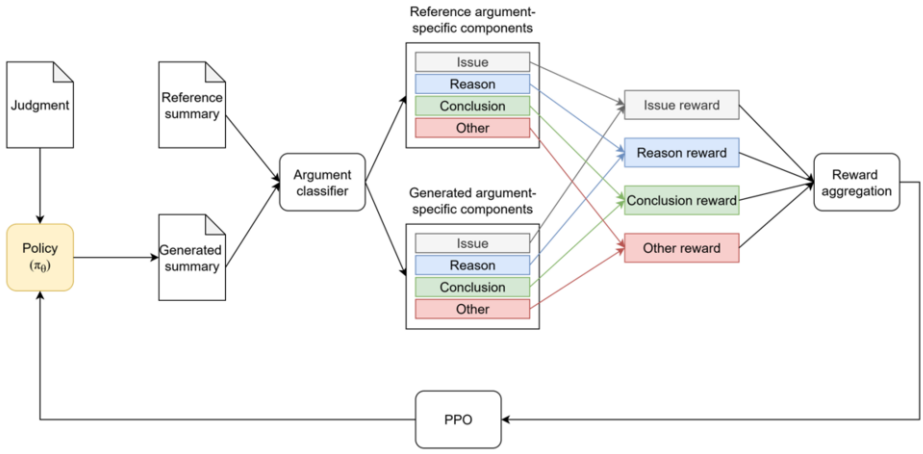



**Figure 1.** Overview structure of our proposed RL-based method. SFT stage is omitted for clarity.

that effectively capture the quality of legal summaries despite their length and structural complexity?

To alleviate the aforementioned problem, we propose an RL-based abstractive legal summarization model with argument-structured reward. Specifically, our method enhances the reward signal with the argumentative structure Issue–Reason–Conclusion (IRC) in legal documents [6]. Our reward design involves three steps: (i) decompose both reference and generated summaries into argument-specific components; (ii) pair the argument-specific components from the reference and generated summaries and compute a sub-reward for each argument-specific component; and (iii) aggregate the sub-rewards according to their importance to obtain the final reward. Intuitively, different argumentative components contribute unequally to the quality of a legal summary, and their importance should be reflected in the reward design. Notably, prior summarization work [7–10] has leveraged argumentative structures in supervised and unsupervised learning, whereas we focus on RL. To the best of our knowledge, this is the first work to incorporate intrinsic argumentative structures into RL-based summarization methods.

We evaluate our argument-structured reward approach with LED and Llama models on the Indian Supreme Court dataset. We compare model performance with and without the argument-structured reward, using BERTScore and ROUGE-L as training rewards. Results demonstrate that our approach achieves consistent improvements over non-argument-structured methods with both models and across both reward settings. With argument-structured reward, Llama achieves gains of up to 4.36, 5.09, 1.62 in ROUGE-1/2/L scores and 0.48 in BERTScore, while LED shows improvements of 3.43, 2.35, 1.83 in ROUGE-1/2/L scores and 0.73 in BERTScore.

## 2. Method

Our proposed method is shown in Figure 1, consisting of three principal steps: (1) policy initialization, (2) argument-structured reward, and (3) policy optimization with argument-structured reward.

### 2.1. Policy Initialization

The policy is initialized by SFT on judgment–summary pairs. Under teacher forcing, SFT minimizes the token-level negative log-likelihood. We then utilize the resulting SFT model to initialize the policy for the RL stage.

### 2.2. Argument-Structured Reward

To obtain more accurate reward for policy optimization, we propose a fine-grained reward mechanism that decomposes summaries into argument-specific components. We compute sub-reward for each component separately and aggregate them into a final reward. Specifically, for each judgment, a summary is generated by the current policy. An argument role classifier assigns one role in $\{I, R, C, O\}$ (Issue, Reason, Conclusion, Other) to each sentence in both reference and generated summaries. For each role in $\{I, R, C, O\}$, we form argument-specific components by concatenating the sentences labeled with this role in the original order. Then we compute the sub-reward for each component pair. For each summary, we thus obtain four sub-rewards. We assign a weight to each sub-reward proportion to the size of the generated section. Intuitively, a role with more tokens is considered more important and receives a larger weight in the total reward. The total reward is the length-weighted sum of sub-rewards.

### 2.3. Policy Optimization with Argument-Structured Reward

Given the total reward, we optimize the policy with Proximal Policy Optimization (PPO) [11]. To prevent the policy from collapsing and keep it close to the SFT reference policy, we consider the Kullback–Leibler (KL) penalty. We form the final reward by combining KL penalty term with the total reward. Finally, we employ PPO to update the policy.

## 3. Argument Role Classifier

Legal summaries usually lack explicit argumentative annotations. To address this issue, we first leverage large language models (LLMs) to automatically generate pseudo-argument-role annotations, and then train an argument role classifier on the LLM-generated data. The trained classifier is subsequently used to assign argument roles to sentences in the summaries.

### 3.1. Multi-LLM Annotation

We employ three LLMs to annotate argument roles at the sentence level. To ensure annotation reliability, we adopt a scheme with following three steps: (i) each summary is split into sentences. The full summary is presented to the annotators so that each decision can be made in the context of the entire document; (ii) Gemini and GPT-4 independently annotate each sentence with one of four roles: Issue, Reason, Conclusion, or Other. They also provide a brief rationale and a confidence score. When two LLMs assign the same role to a sentence, this role is directly adopted as the final annotation for that sentence; (iii) when Gemini and GPT-4 disagree on a sentence, their rationales and confidence

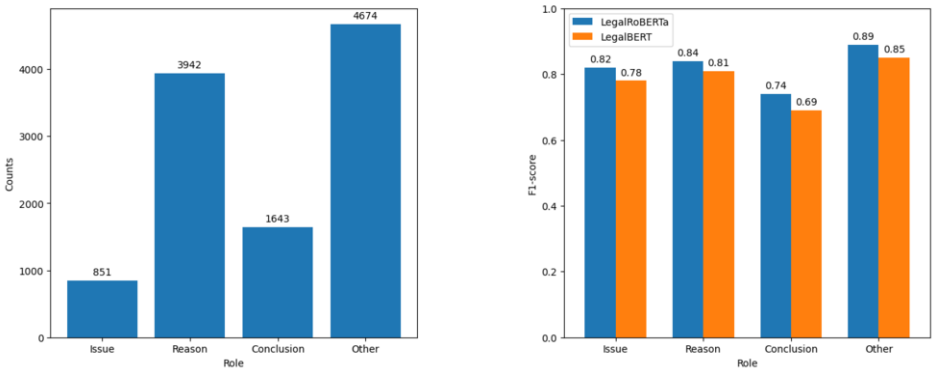



**Figure 2.** Annotation results (left) and classifier performance (F1-scores) on the test dataset (right).

scores, together with the sentence and full summary, are provided to Claude. Acting as an arbitrator, Claude determines the final annotation with rationale and confidence.

We apply our annotation strategy to the Indian Supreme Court dataset [12]. In total, 695 cases are annotated. We use 595 cases for training the classifier and 100 cases for testing. Figure 2 (left) shows the annotation results. Gemini and GPT-4 agreed on 9,622 sentences, yielding an agreement rate of 86.6%. The Cohen's $\kappa$ averaged across all annotations is 0.81, which indicates substantial agreement.

### *3.2. Classifier Performance*

Based on the pseudo-annotated dataset, we train LegalBERT and LegalRoBERTa [13] classifiers for four-way classification. Figure 2 (right) shows the results on the test dataset. LegalRoBERTa is selected as the backbone classifier in our main experiments due to its stronger and more stable performance.

## 4. Experiment Setting

We conduct experiments on an abstractive legal summarization dataset: Indian Supreme Court dataset [12]. We select Longformer-Encoder-Decoder (LED) [14] as our encoder–decoder baseline and Llama-3.1-8B [15] as our LLM baseline. We evaluate generation quality with ROUGE and BERTScore , which are widely used in abstractive summarization benchmarks. We report ROUGE-1, ROUGE-2, and ROUGE-L F1 score. For BERTScore, we use the pretrained model roberta-large and report the F1 score.

## 5. Results

Table 1 shows overall results. We employ either ROUGE-L or BERTScore as the reward in RL and consider both argument structured and unstructured methods to construct reward. We observe that RL *without* structured reward is unstable, whereas RL *with* structured reward achieves promising improvement which is *metric-specific*. Namely, when a particular metric is used as the reward, cross-metric improvement is limited.

| | | | ROUGE-1 | ROUGE-2 | ROUGE-L | BERTScore |
|---|---|---|---|---|---|---|
| LED | SFT | baseline | 49.89 | 25.91 | 27.62 | 83.06 |
| | RL | ROUGE-L *w/o* stru | 49.91 | 25.84 | 27.63 | 83.07 |
| | | ROUGE-L *w/*stru | **53.32** | **28.26** | **29.45** | 83.09 |
| | | BERTScore *w/o* stru | 49.85 | 25.81 | 27.71 | 83.07 |
| | | BERTScore *w/*stru | 49.87 | 25.92 | 27.61 | **83.79** |
| Llama | SFT | baseline | 54.32 | 33.12 | 35.42 | 83.04 |
| | RL | ROUGE-L *w/o* stru | 54.38 | 33.61 | 35.68 | 83.01 |
| | | ROUGE-L *w/*stru | **58.68** | **38.21** | **37.04** | 83.03 |
| | | BERTScore *w/o* stru | 54.29 | 33.1 | 35.41 | 83.06 |
| | | BERTScore *w/*stru | 54.3 | 33.14 | 35.44 | **83.52** |

**Table 1.** Evaluation results of LED and Llama on the test dataset. '*w/o* stru' indicates RL training *without* structured reward; '*w/* stru' indicates RL training *with* structured reward. The highest scores are in bold.

LED results indicate that when RL is applied with unstructured rewards (either ROUGE-L or BERTScore), the improvement over the SFT baseline is marginal, and performance on some metrics even decreases. This differs from common observations on conventional news summarization where setting a target metric as reward often directly improves that metric (*metric-specific improvement*). By contrast, RL with structured rewards yields stable improvements. Specifically, under structured ROUGE-L reward, ROUGE-1/2/L increase by 3.43, 2.35 and 1.83, respectively; under structured BERTScore reward, BERTScore increases by 0.73. Notably, structured ROUGE-L does not meaningfully raise BERTScore, and structured BERTScore does not meaningfully raise ROUGE scores or even could reduce them. Llama results illustrate the same pattern. Compared with its SFT baseline, RL with structured ROUGE-L reward yields gains on ROUGE-1/2/L by 4.36, 5.09 and 1.62, respectively, while BERTScore remains essentially unchanged; structured BERTScore reward improves BERTScore by 0.48, whereas ROUGE is nearly unchanged.

## 6. Conclusion and Future Work

We proposed an argument-structured reward framework that leverages the intrinsic IRC structure of legal documents to provide more fine-grained and reliable optimization signals. Experiments on the Indian Supreme Court dataset with both LED and Llama show consistent improvements across reward settings. In future work, we plan to investigate training the classifier on human-annotated datasets to further improve the performance.

## Acknowledgements

This work was supported by the "R&D Hub Aimed at Ensuring Transparency and Reliability of Generative AI Models" project of the Ministry of Education, Culture, Sports, Science and Technology and JSPS KAKENHI Grant Numbers, JP22H00543, JP25H00522, JP25H01112 and JP25H01152.

## References


[1] Ryu S, Do H, Kim Y, Lee G, Ok J. Multi-Dimensional Optimization for Text Summarization via Reinforcement Learning. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics; 2024. p. 5858-71.

[2] Parnell J, Jauregi Unanue I, Piccardi M. A Multi-Document Coverage Reward for RELAXed Multi-Document Summarization. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics; 2022. p. 5112-28.

[3] Parnell J, Jauregi Unanue I, Piccardi M. RewardsOfSum: Exploring Reinforcement Learning Rewards for Summarisation. In: Proceedings of the 5th Workshop on Structured Prediction for NLP; 2021. p. 1-11.

[4] Krishna K, Bransom E, Kuehl B, Iyyer M, Dasigi P, Cohan A, et al. LongEval: Guidelines for Human Evaluation of Faithfulness in Long-form Summarization. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics; 2023. p. 1650-69.

[5] Guo X, Vosoughi S. Length Does Matter: Summary Length can Bias Summarization Metrics. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; 2023. p. 15869-79.

[6] Xu H, Šavelka J, Ashley KD. Using argument mining for legal text summarization. In: Legal knowledge and information systems; 2020. p. 184-93.

[7] Elaraby M, Litman D. ArgLegalSumm: Improving Abstractive Summarization of Legal Documents with Argument Mining. In: Proceedings of the 29th International Conference on Computational Linguistics; 2022. p. 6187-94.

[8] Elaraby M, Zhong Y, Litman D. Towards Argument-Aware Abstractive Summarization of Long Legal Opinions with Summary Reranking. In: Findings of the Association for Computational Linguistics: ACL 2023; 2023. p. 7601-12.

[9] Zhong Y, Litman D. STRONG–Structure Controllable Legal Opinion Summary Generation. arXiv preprint arXiv:230917280. 2023.

[10] Zhong Y, Litman D. Computing and Exploiting Document Structure to Improve Unsupervised Extractive Summarization of Legal Case Decisions. In: Proceedings of the Natural Legal Language Processing Workshop 2022; 2022. p. 322-37.

[11] Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. arXiv preprint arXiv:170706347. 2017.

[12] Shukla A, Bhattacharya P, Poddar S, Mukherjee R, Ghosh K, Goyal P, et al. Legal Case Document Summarization: Extractive and Abstractive Methods and their Evaluation. In: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing); 2022. p. 1048-64.

[13] Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I. LEGAL-BERT: The muppets straight out of law school. arXiv preprint arXiv:201002559. 2020.

[14] Beltagy I, Peters ME, Cohan A. Longformer: The long-document transformer. arXiv preprint arXiv:200405150. 2020.

[15] Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:230213971. 2023.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251604*

# A Hohfeldian Knowledge Base for LLM-Assisted Legal Information Retrieval in Marine Biodiversity Law

Rohan NANDA [a,b,1] Henrique MARCOS [c] Hannes WESTERMANN [a]
Julia SCHUTZ VEIGA [d]

[a] *Law and Tech Lab, Faculty of Law, Maastricht University*
[b] *Department of Advanced Computing Sciences, Maastricht University*
[c] *Faculty of Law, Maastricht University; CEDMAR, So Paulo University*
[d] *NOVA School of Law; Ocean Voices, Nippon Foundation, University of Edinburgh*
ORCiD ID: Rohan Nanda https://orcid.org/0000-0001-7124-4799, Henrique Marcos https://orcid.org/0000-0003-0259-5530, Hannes Westermann https://orcid.org/0000-0002-4527-7316, Julia Schutz Veiga https://orcid.org/0000-0002-2504-7353

**Abstract.** Governance of marine genetic resources is fragmented across overlapping treaties, creating uncertainty about which obligations apply in specific situations. We address this by mapping treaty provisions into a normative structure based on Hohfelds framework, the Hohfeld-Structured Normative Knowledge Base (HSNKB), and by constructing a dataset of 15 fact-pattern questions with expert gold answers. We evaluate four recent large language models (LLMs) on retrieving rows that contain the relevant rules to answer the fact-pattern questions. The results indicate that reasoning LLMs achieved modest precision and middling recall. Hohfeldian representations help avoid false positives, but improving recall without degrading precision remains an open problem for cross-treaty retrieval.

**Keywords.** Hohfeld, legal reasoning, legal information retrieval, large language models (LLMs)

## 1. Introduction

Global biodiversity governance is spread across overlapping international agreements, resulting in a complex patchwork of legal obligations for states and other actors [1]. This makes it difficult for governments and stakeholders to discern who holds which duties or entitlements in a given scenario, especially where new obligations, such as the ones introduced by the Biodiversity Beyond National Jurisdiction Agreement (BBNJ-A), intersect with preexisting benefit-sharing rules under the Convention on Biological Diversity (CBD) and the Nagoya Protocol (NP). To address this challenge, we introduce a Hohfeld-Structured Normative Knowledge Base (HSNKB) that unifies key provisions from the

[1] Corresponding Author: Rohan Nanda, r.nanda@maastrichtuniversity.nl.

CBD, NP, and BBNJ-A. Our approach is grounded in Hohfelds theory [2] of legal positions, which provides a fine-grained framework to formalise normative relationships. Our work offers an initial evaluation of how LLMs perform cross-treaty retrieval using structured norms, aiming to help legal and policy actors assess overlapping obligations.

The major contributions of this paper are as follows: (1) **Two datasets:** (a) HSNKB: covering three international treaties: CBD, NP, and BBNJ-A; and (b) a benchmark dataset of 15 fact-pattern legal questionanswer pairs, where each gold-standard answer consists of the relevant HSNKB rows from the three treaties that resolve the fact-pattern scenario. (2) **Retrieval pipelines:** We develop two legal information retrieval pipelines (full-scope and actor-constrained) that leverage the HSNKB to answer the fact-pattern questions by selecting the relevant rows. Both pipelines use a zero-shot Chain-of-Instructions [3] prompt and structured JSON output. (3) **Comparative Evaluation:** We benchmark the fact-pattern dataset with four state-of-the-art LLMs, comparing the predicted sets of rows to the expert gold sets for each question using macro-average precision, recall, and F1. Our research questions are: **RQ1** : How accurately can LLMs retrieve the relevant rows from the HSNKB when answering fact-pattern legal questions across international treaties? **RQ2** : How well do LLMs identify the actors that are relevant for a specific fact pattern? **RQ3** : Does actor-constrained retrieval improve performance compared to the full-scope retrieval?

We propose a system that reads a complex fact-pattern question, ingests the HSNKB (see section 3) of international biodiversity instruments, identifies relevant provisions across instruments, and visualises the resulting actor relationships (see Figure 1), thus supporting more transparent legal reasoning and access to justice.

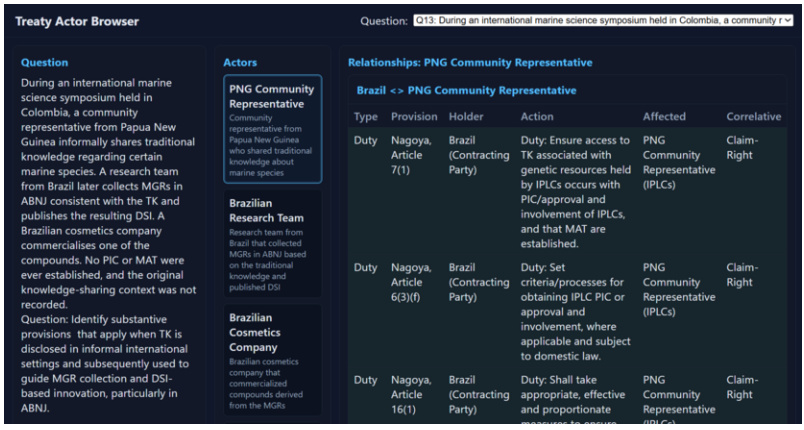



**Figure 1.** Prototype interface showing the analysed question, relevant actors, and their treaty relationships

## 2. Related Work

The use of Hohfelds analytical framework in legal informatics has long been motivated by the difficulty of expressing normative concepts without collapsing them into vague categories. Markovich [4] shows how Hohfelds schema can be formalised to preserve the relational and directed character of legal positions, arguing that precise differenti-

ation of deontic types is essential if computational models are to capture the logic of legal relations without distortion. Francesconi [5], in turn, develops a semantic model that annotates and reasons over normative provisions, introducing patterns for Hohfeldian relations and enabling both explicit and implicit norms to be retrieved across legislative texts. Peters and Wyner [6] develop a semi-automatic workflow that uses NLP tools and expert feedback to extract Hohfeldian relations from legal texts. Our work also connects to legal question answering, a key use case of artificial intelligence and law (for early work, see, e.g., [7,8,9]). More recently, LLMs have been utilised to support legal question-answering. Approaches include using such models to support legal expert systems (e.g., [10,11] asking a model to answer legal questions directly (see, e.g., [12,13,14]), or connecting LLMs to more traditional information retrieval systems, e.g., in a retrieval-augmented generation setup (see e.g., [15,16,17]).

## 3. Dataset

**Dataset 1: Hohfeld-Structured Normative Knowledge Base (HSNKB):** Two international law experts constructed the HSNKB from three core international instruments governing biodiversity (CBD, NP, BBNJ-A). These treaties were selected because they establish the principal normative frameworks for access and benefit-sharing (ABS) of genetic resources. Each row in the HSNKB represents a single operative clause (article, paragraph, or sub-paragraph) expressing one dominant normative relation between actors. Provisions containing multiple prescriptions were divided into separate entries, while preambular text, aspirational language, and definitional provisions were excluded. Each entry includes a unique identifier, treaty name, provision number, the actor-holder, the legal position (or action), the actor-affected, and an expert-written natural-language sentence describing the tripartite relation between actor-holder, legal position and actor-affected. Correlative positions are deduced based on Hohfeldian logic. HSNKB contains 377 rows (193 from BBNJ-A, 96 from CBD, and 88 from NP), with "Contracting Parties" as the most frequent actor and "duty" as the dominant Hohfeldian relation.

**Dataset 2: Fact Pattern Questions & Relevant Rows:** We developed 15 fact-pattern questions that connect the HSNKB to doctrinal reasoning for the legal information retrieval task. The relevant rows from the HSNKB that answer each question have been annotated as ground truth by an international law expert in marine biodiversity. Each question is formulated according to the *IRAC* method: a scenario that presents actors and circumstances, followed by an implicit issue, inviting identification of the applicable rule and analysis of its consequences. This format moves from straightforward single-treaty scenarios to more complex cases that require cross-referencing across multiple instruments. The dataset size is purposefully small to enable careful expert annotation. We plan to expand its coverage and inter-annotator validation in future work.

**Actor Definitions** We compiled a dictionary of actor definitions linking names to descriptions. This standardises references to implicit actors in fact-patterns and ensures consistent mapping across instruments.

## 4. Experimental design

To evaluate how well LLMs retrieve relevant treaty provisions from the HSNKB to answer fact-pattern questions, we implemented two retrieval pipelines: full-scope retrieval and actor-constrained retrieval. In the full-scope retrieval, the model receives all HSNKB rows with three columns (unique identifier, provision number and the natural language sentence conveying the tripartite relation), along with supporting definitions (actor definitions, explanation of Hohfeldian positions) embedded in the prompt and is asked to identify the treaty provisions relevant to each fact pattern. In actor-constrained retrieval, the model first identifies the relevant actors in a fact-pattern, filters the HSNKB to candidate rows involving those actors, and then retrieves the relevant rows from this reduced set. Both pipelines follow a zero-shot Chain-of-Instructions (CoI) [3] strategy. The system prompt instructions guide the model through four main stages: (i) parsing key entities in the question such as actors, jurisdiction, resource type; (ii) applying jurisdictional gating to avoid mixing provisions across regimes; (iii) applying an allocative filter to retain only provisions conferring Hohfeldian positions to the relevant actors; and (iv) extracting HSNKB rows ranked by their specificity to jurisdiction, resource type, lifecycle stage, and traditional knowledge. For each question, the LLM is then asked to return the unique identifiers of the relevant rows in structured JSON format.

The pipeline is run with four different LLMs (Gemini 2.5 Pro, GPT-5, Claude Sonnet 4, and Kimi K2) accessed via the OpenRouter[2] API. All models were run in a zero-shot setting, with temperature set to 0 (except GPT-5 which does not support the temperature parameter), a fixed seed, and response format as JSON. For the three reasoning models (Gemini 2.5 Pro, GPT-5, Claude Sonnet 4), the effort parameter was set to high so that the model allocates a large portion of tokens for internal reasoning. The only expected output is a dictionary of ranked unique identifiers corresponding to HSNKB rows. For each of the 15 fact-pattern questions, the predicted unique identifiers (for rows) by the LLM are compared against the gold-standard rows. We utilise standard information retrieval metrics: precision, recall, and F1-score. Macro-averaged precision, recall, and F1-scores are computed across all 15 questions, ensuring each question contributes equally to the overall results. The dataset, code, and prompt are available at our GitHub repository[3].

## 5. Results and Discussion

This section presents the results of our evaluation with respect to the research questions. **RQ1** examines how accurately LLMs retrieve the relevant rows from the HSNKB. In the full-scope retrieval (see Table 1), GPT-5 attains the best performance, indicating strong selectivity with moderate recall. The three reasoning LLMs: GPT-5, Gemini 2.5 Pro, and Claude Sonnet 4 outperform the only non-reasoning LLM, Kimi K2. Across the two best reasoning models, precision was modest, while recall was middling. Therefore, the main challenge is improving recall without reducing precision. Error analysis indicates that false negatives mainly involved conditional or multi-actor duties dispersed across treaties, particularly those linked to benefit-sharing obligations. False positives were less frequent and stemmed from the inclusion of programmatic cooperation clauses which go beyond what was required by the fact-pattern.

[2] https://openrouter.ai/

[3] https://github.com/rnanda17/jurix25-hohfledian-KB-LLM-Assisted-Legal-IR/

**Table 1.** Macro-average precision (P), recall (R), and F1-Score: Full-Scope vs Actor-Constrained Retrieval

| | Full-Scope Retrieval | | | Actor-Constrained Retrieval | | |
|---|---|---|---|---|---|---|
| Model | P | R | F1 | P | R | F1 |
| Gemini 2.5 Pro | 0.69 | 0.45 | 0.53 | 0.68 | 0.49 | 0.56 |
| Claude Sonnet 4 | 0.61 | 0.50 | 0.52 | 0.64 | 0.44 | 0.51 |
| GPT-5 | 0.77 | 0.53 | **0.59** | 0.74 | 0.50 | 0.57 |
| Kimi K2 | 0.28 | 0.58 | 0.33 | 0.30 | 0.58 | 0.35 |

**RQ2** addresses how well LLMs identify the actors relevant to each fact-pattern question. Table 2 shows consistently high recall across all models, confirming that most relevant actors were recognized. On average, the model predicted between 24 and 36 actors per question, which is significantly higher than the gold standard, where only 6.2 actors are retrieved. This outcome reflects the instruction to be inclusive for actor selection and the fact that models identified both holders and affected parties. GPT-5 and Claude Sonnet performed well. Kimi K2, the only non-reasoning model, performed slightly worse.

**Table 2.** Average Precision, Recall, F1-score for the retrieved actors and average number of actors retrieved.

| **Model** | **Precision** | **Recall** | **F1-score** | **Actors Retrieved** |
|---|---|---|---|---|
| Gemini 2.5 Pro | 0.171 | 0.944 | 0.286 | 35.6 |
| Claude Sonnet 4 | 0.234 | 0.992 | 0.372 | 27.7 |
| GPT-5 | 0.239 | 1.000 | 0.379 | 27.7 |
| Kimi K2 | 0.238 | 0.880 | 0.368 | 23.9 |

**RQ3** compares the full-scope and actor-constrained retrieval pipelines. The results (Table 1) show only limited differences relative to the full-scope retrieval. Actor-constrained filtering narrows the search space but does not substantially improve retrieval performance. In general, only around 15% data was filtered. The reason is that most rows have "Contracting Party" as the actor-holder and it was present in all fact-patterns. Thus, most rows had to be included as candidate rows, limiting the potential gains of filtering.

## 6. Conclusion

This paper presented HSNKB, a fact-pattern dataset, and an evaluation of four LLMs using two retrieval pipelines. We compared two retrieval pipelines: full-scope retrieval and actor-constrained retrieval. The reasoning LLMs achieved modest precision and middling recall. Actor-constrained retrieval yields limited gains over full-scope retrieval. Improving recall while keeping precision stable is therefore an important next step for cross-treaty retrieval. The error analysis of retrieval performance suggests that false negatives cluster around conditional and multi-actor duties, while occasional false positives arise when cooperation language was interpreted as Hohfeldian duties. Future work will extend the dataset with more fact-patterns and include information retrieval baselines to better contextualize LLM-assisted retrieval.

## Acknowledgements

This research was supported by a grant from the Netherlands ELS Academy.

## References


[1] Morgera E. Fair and equitable benefit-sharing in a new international instrument on marine biodiversity: A principled approach towards partnership building. Maritime Safety and Security Law Journal. 2018;5:66.

[2] Hohfeld WN. Some Fundamental Legal Conceptions as Applied in Judicial Reasoning. The Yale Law Journal. 1913;23(1):16-59.

[3] Zin MM, Satoh K, Borges G. Leveraging LLM for Identification and Extraction of Normative Statements. In: Legal Knowledge and Information Systems. IOS Press; 2024. p. 215-25.

[4] Markovich R. Understanding Hohfeld and Formalizing Legal Rights: The Hohfeldian Conceptions and Their Conditional Consequences. Studia Logica. 2020;108(1):129-58.

[5] Francesconi E. Semantic Model for Legal Resources: Annotation and Reasoning over Normative Provisions. Semantic Web. 2016 Mar;7(3):255-65.

[6] Peters W, Wyner A. Legal Text Interpretation: Identifying Hohfeldian Relations from Text. In: Calzolari N, Choukri K, Declerck T, Goggi S, Grobelnik M, Maegaard B, et al., editors. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). Portorož, Slovenia: European Language Resources Association (ELRA); 2016. p. 379-84. Available from: https://aclanthology.org/L16-1059/.

[7] Sergot MJ, Sadri F, Kowalski RA, Kriwaczek F, Hammond P, Cory HT. The British Nationality Act as a logic program. Communications of the ACM. 1986;29(5):370-86.

[8] Susskind R. Expert systems in law. Oxford University Press, Inc.; 1987.

[9] Westermann H, Benyekhlef K. JusticeBot: A Methodology for Building Augmented Intelligence Tools for Laypeople to Increase Access to Justice. In: Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law. ICAIL '23. New York, NY, USA: Association for Computing Machinery; 2023. p. 351-60. Available from: https://doi.org/10.1145/3594536.3595166.

[10] Janatian S, Westermann H, Tan J, Savelka J, Benyekhlef K. From text to structure: Using large language models to support the development of legal expert systems. In: Legal Knowledge and Information Systems. IOS Press; 2023. p. 167-76.

[11] Westermann H, et al. Bridging the gap: Mapping layperson narratives to legal issues with language models. In: ASAIL 2023; 2023. .

[12] Tan J, Westermann H, Benyekhlef K. ChatGPT as an Artificial Lawyer? In: AI4AJ 2023; 2023. .

[13] Dahl M, Magesh V, Suzgun M, Ho DE. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. Journal of Legal Analysis. 2024 Jan;16(1):6493. Available from: http://dx.doi.org/10.1093/jla/laae003.

[14] Marcos H. Can large language models apply the law? AI & Society. 2025;40(5):3605-14.

[15] Savelka J, Ashley KD, Gray MA, Westermann H, Xu H. Explaining Legal Concepts with Augmented Large Language Models (GPT-4). arXiv preprint arXiv:230609525. 2023.

[16] Louis A, Dijck Gv, Spanakis G. Interpretable Long-Form Legal Question Answering with Retrieval-Augmented Large Language Models. Proceedings of the AAAI Conference on Artificial Intelligence. 2024 Mar;38(2020):22266–22275.

[17] Pipitone N, Alami GH. LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain. 2024 Aug;(arXiv:2408.10343). Available from: http://arxiv.org/abs/2408.10343.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251605*

# Comparison of Embedding Methods for Retrieval Under Noisy Institutional Labels

Tereza NOVOTNÁ [a,1], and Jakub HARAŠTA [a]
[a] *Faculty of Law, Masaryk University, Brno, Czechia*

**Abstract.** Retrieving relevant case law remains a time-consuming task. We compare two embedding models for Czech Constitutional Court decisions: (i) a large general-purpose OpenAI embedder and (ii) a domain-specific BERT trained from scratch on ∼34,000 decisions. We introduce a noise-aware evaluation using IDF-weighted keyword overlap as graded relevance, dual thresholds (0.20, 0.28), paired-bootstrap significance, and nDCG diagnostics. Despite conservative absolute nDCG due to noisy institutional labels, the OpenAI embedder consistently and significantly outperforms the domain BERT across all ranks and thresholds. Our framework enables robust evaluation under imperfect gold standards typical of legacy judicial databases.

**Keywords.** legal information retrieval, case law, embeddings, evaluation, noisy labels, Czech Constitutional Court

## 1. Introduction and Background

Searching for similar decisions constitutes a large part of the work of analysts, clerks, and judges at the Czech Constitutional Court (ConC). We evaluate the feasibility of deploying automated methods to support the retrieval of semantically similar decisions.

ConC provides public access to its case law via the NALUS database.[2] Users typically retrieve cases by legal provisions or keywords. Currently, 713 keywords are manually assigned to decisions by ConC staff, making the set a gold standard. However, staff turnover, the absence of formal rules, and emerging legal issues introduce noise and inconsistency over time. This problem is well known in Czechia, but its impact, including for automation, remains understudied.

We compare a domain-trained BERT with a general OpenAI embedder across three embedding-based retrieval methods and propose an evaluation tolerant to noisy gold: IDF-weighted keyword overlap, two data-driven relevance thresholds (0.20, 0.28), paired bootstrap, and nDCG diagnostics. Results show the general embedder clearly outperforms domain BERT, offering a reproducible framework for semantic retrieval under noisy institutional metadata.

Text embeddings and TLMs/LLMs [1, 2, 3] are regularly implemented to enhance document retrieval. Pre-trained Legal-BERT [4] contributes to increased performance for domain-specific tasks. In case law retrieval, efforts include unsupervised techniques [5]

[1] Corresponding author: Tereza Novotná, email: tereza.novotna@law.muni.cz.
[2] nalus.usoud.cz

or inclusion of supporting tools and models into the retrieval pipeline [6, 7, 8, 9]. We retrieve ConC case law using text embeddings and evaluate them.

LLMs add value in domain-specific tasks via zero-/few-shot capabilities [10,11,12], but evaluation is hindered by scarce legal benchmarks [13]. Recent comparative efforts [14] inform our approach: we evaluate the OpenAI embedder through robust benchmarking on established legal tasks.

## 2. Data and Benchmark

### 2.1. Corpus and Metadata

By combining a subset of ConC decisions from the Czech Court Decisions Corpus (1993-2018) (CzCDC) [3] with retrieval from NALUS (2018-2024), we obtained all the ConC's decisions between its establishment in January 1993 and December 2024, including their metadata (e.g., keywords). We further used only the 34,441 decisions longer than 10,000 characters. [4]

A typical decision has three keywords (entropy $\approx$ 6.9 bits, effective vocabulary $\approx$ 120), with 0.08 expected overlap and 93% zero-overlap between similar cases. In more diverse years (entropy $\approx$ 7.8, $N_{\text{eff}} \approx 230$), overlap halves and zero-overlap reaches to 96%. Even with five keywords, overlap remains low (0.21; 81% zero). Over time, growing label diversity kept overlap-based metrics conservative while preserving relative model differences.

We form a query set of 2,000 decisions from the dataset, each with assigned keywords (metadata). To ensure that keyword overlap is at least measurable, we restrict the sampling pool to decisions with $\geq 2$ keywords. We then apply stratified sampling along three axes that affect evaluation difficulty: (1) keyword count per decision, binned as 2–3, 4–7, 8+, (2) rarity of a decision's tags, computed as the *median* IDF of its (deduplicated) keyword set (IDF defined below) and split at the median into *common* vs. *rarer*, (3) decision year binned by empirical quantiles (three bins; ties handled by `qcut` with duplicates allowed). Within the resulting $3 \times 2 \times 3 = 18$ strata, we allocate samples approximately evenly with a fixed random seed (reallocating surplus to larger strata when a stratum is under-populated).

### 2.2. Graded Relevance via IDF-Weighted Overlap

We measure similarity between two decisions using an IDF-weighted Jaccard overlap (*wJacc*), where rarer keywords contribute more to the score.

The resulting *wJacc* value serves as a graded relevance score ($r_i$) for computing nDCG. Binary metrics such as P@$k$, MAP@$k$, and RBP are derived by thresholding $r_i$ at two empirically calibrated levels of strictness, denoted by $\tau$ (0.20 for balanced and 0.28 for strict evaluation). The balanced threshold captures moderate overlaps, while the strict one isolates high-confidence matches based on percentile distributions of *wJacc* among retrieved and ideal candidates. Recall@$k$ is omitted, as the large number of

[3] hdl.handle.net/11372/LRT-3052

[4] Proxy to distinguish between procedural decisions lacking substantive legal argumentation, and substantive decisions.

$\tau$-relevant items makes it less informative; instead, graded nDCG and head-oriented P@$k$/HitRate@$k$/RBP better reflect user utility under noisy, heterogeneous labels.

## 3. Methods

### 3.1. Models

**BERT Pre-trained from scratch on ConC corpus.** We pretrained a BERT-style encoder from scratch on ∼34k ConC decisions. Markup and boilerplate were removed while preserving punctuation, numerals, and Czech diacritics. Text was normalized to NFC and kept *cased*. A 32k subword vocabulary was learned using SentencePiece (Unigram)[5], which accommodates Czech morphology. The model follows the BERT-base architecture (12 layers, hidden size 768, 12 attention heads, GELU activations, dropout 0.1) and is trained from random initialization without multilingual pretraining. To address the 512-token limit, documents were split into overlapping 512-token windows (stride $s$=256). Pretraining used masked language modeling (MLM) only, with 15% of tokens masked under the 80/10/10 rule and dynamic masking per epoch. Training ran for 10 epochs on a single GPU.

Each window produces hidden states $\{h_1, \ldots, h_T\}$ (excluding [CLS], [SEP]). We evaluated two pooling schemes: (1) mean pooling: $e_{\text{win}} = \frac{1}{T}\sum_{i=1}^{T} h_i$, (2) self-attention pooling: $\alpha_i \propto \exp(w^\top h_i)$, $e_{\text{win}} = \sum_i \alpha_i h_i$, where $w$ is a learnable vector. Document embeddings are obtained by averaging window embeddings: $e_{\text{doc}} = \frac{1}{|W|}\sum_{w \in W} e_{\text{win}}^{(w)}$, followed by L2 normalization for cosine retrieval. We report results for the attention-based variant, which consistently outperformed mean pooling. IR relies solely on cosine similarity in the learned embedding space.

**OpenAI (general-purpose) model.** We use OpenAI's `text-embedding-3-large` as an off-the-shelf embedder. The model produces 3,072-dimensional vectors by default and is the current high-accuracy embedding family from OpenAI, with documented gains over prior `ada-002` on MIRACL and MTEB benchmarks. It also supports an optional *dimensions* parameter to shorten embeddings (for storage/speed) while retaining most retrieval quality.

### 3.2. Document Scope Settings, Retrieval Setup

We build two BERT variants differing in input scope: (*i*) **Full decisions**, split into 512-token windows and averaged into one vector, and (*ii*) **Facts only**, embedding just the factual background extracted by simple layout rules. The latter tests whether shorter, denser sections reduce *topic-averaging* noise. The OpenAI embedder also uses the *facts-only* scope for comparability and lower token cost.

All embeddings are L2-normalized, reducing cosine similarity to the inner product $\cos(e_q, e_d) = e_q^\top e_d$. Each decision serves as a query retrieving its top-$k$ neighbours ($k \in \{10, 20, 100\}$) excluding self-matches. Retrieval uses exact inner-product kNN with FAISS `IndexFlatIP` (float32). No lexical filters or re-ranking are applied. For fairness, the query set is the intersection of documents across models, with deterministic tie-breaking and fixed seeds ensuring reproducibility.

[5] We preserve diacritics and avoid heuristic splitting beyond SentencePiece rules.

**Table 1.** Results at $k \in \{10, 20, 100\}$ with graded nDCG and binary metrics at $\tau \in \{0.20, 0.28\}$ over 2,000 queries. nDCG is $\tau$-independent ($r = \text{wJacc}$).

| | $k = 10$ | | $k = 20$ | | $k = 100$ | | Weighted/Cnt@10 |
|---|---|---|---|---|---|---|---|
| Model | nDCG | P/Hit (.20/.28) | nDCG | P/Hit (.20/.28) | nDCG | P/Hit (.20/.28) | (approx.) |
| OpenAI-embedder_facts | **0.190** | **0.260/0.716–0.168/0.550** | **0.187** | **0.236/0.802–0.151/0.649** | **0.182** | **0.179/0.936–0.109/0.827** | **42.8 / 12.2** |
| BERT_ab_all | 0.090 | 0.106/0.467–0.060/0.315 | 0.090 | 0.096/0.600–0.053/0.406 | 0.090 | 0.070/0.855–0.038/0.657 | 27.8 / 8.6 |
| BERT_ab_facts | 0.060 | 0.067/0.364–0.038/0.231 | 0.059 | 0.058/0.472–0.032/0.309 | 0.061 | 0.045/0.774–0.024/0.571 | 14.3 / 4.8 |

### *3.3. Evaluation Metrics and Significance*

We evaluate ranked retrieval using graded and binary relevance. Each decision is a query, with neighbours scored by IDF-weighted Jaccard overlap (*wJacc*) of their keywords. Graded relevance uses this score directly, whereas binary metrics apply two thresholds (0.20 and 0.28). The *nDCG@k* metric captures both relevance strength and rank position, remaining $\tau$-independent and robust under noisy labels.

Binary metrics include *Precision@k* (the share of relevant items among the top $k$) and *Mean Average Precision* (*MAP@k*), which reward ranking relevant cases early. *HitRate@k* measures the likelihood of retrieving at least one relevant item ("first useful result"), while *Rank-Biased Precision* (*RBP@10*) models user patience by emphasizing early ranks. For diagnostics, *WeightedOverlap@k* and *OverlapCount@k* quantify how many—and how distinctive—keywords are shared between query and retrieved cases, clarifying nDCG trends under diverse label distributions. Statistical significance is assessed via paired bootstrap resampling over queries, reporting mean differences, 95% confidence intervals, and two-sided $p$-values.

## 4. Results

Across all settings ($k \in \{10, 20, 100\}$, $\tau \in \{0.20, 0.28\}$), the OpenAI embedder consistently outperforms both domain-specific BERT variants in nDCG, P@k, and HitRate@k (statistically significant under paired bootstrap tests). We binarize graded gains $r = \text{wJacc}(A, B)$ at thresholds $\tau \in \{0.20, 0.28\}$ to obtain precision-style metrics; as $\tau$ increases, relevance becomes more stringent and P@k and HitRate@k drop accordingly, while relative model ordering remains stable. At $k$=10, for instance, P@10 decreases from 0.260 to 0.168 and HitRate@10 from 0.716 to 0.550 as $\tau$ rises from 0.20 to 0.28, whereas graded nDCG@10 stays constant at 0.190. Additional metrics follow the same pattern: MAP improves from 0.0008 (BERT_ab_all) and 0.0003 (BERT_ab_facts) to 0.0029–0.0050 for the OpenAI embedder, and RBP@10 rises from 0.071 and 0.046 to 0.175–0.113 across thresholds. These consistent gains confirm that the OpenAI embedder retrieves more semantically aligned neighbours and ranks them earlier, reinforcing its robust advantage across all evaluation criteria.

All head-to-head gaps at $k = 10$ are statistically significant under paired bootstrap (Table 2), with tight 95% confidence intervals across both $\tau$ tracks: OpenAI outperforms Bert models on nDCG@10, P@10, and HitRate@10 with $p < 0.001$ in every case. Together with our label-drift analysis, this supports the conclusion that absolute nDCG is conservative under diversified labels, yet relative differences are large and statistically significant.

**Table 2.** Paired bootstrap over 2,000 common queries ($B = 2000$ resamples). Δ = mean difference A−B; 95% CI in brackets. All $p < 0.001$.

| | | $\tau = 0.20$ | | | $\tau = 0.28$ | | |
|---|---|---|---|---|---|---|---|
| Comparison | Metric | Δ | 95% CI | $p$ | Δ | 95% CI | $p$ |
| *OpenAI-embedder_facts − BERT_ab_all* | | | | | | | |
| | nDCG@10 | **+0.099** | [0.093, 0.105] | <0.001 | **+0.099** | [0.093, 0.105] | <0.001 |
| | P@10 | **+0.154** | [0.144, 0.165] | <0.001 | **+0.108** | [0.099, 0.116] | <0.001 |
| | Hit@10 | **+0.250** | [0.224, 0.275] | <0.001 | **+0.236** | [0.211, 0.262] | <0.001 |
| *OpenAI-embedder_facts − BERT_ab_facts* | | | | | | | |
| | nDCG@10 | **+0.130** | [0.124, 0.136] | <0.001 | **+0.130** | [0.124, 0.136] | <0.001 |
| | P@10 | **+0.194** | [0.183, 0.204] | <0.001 | **+0.130** | [0.121, 0.139] | <0.001 |
| | Hit@10 | **+0.353** | [0.329, 0.378] | <0.001 | **+0.319** | [0.296, 0.343] | <0.001 |

## 5. Discussion and Conclusion

Our from-scratch BERT lags behind the off-the-shelf OpenAI embedder. The likely reason is scale and objective: the general embedder benefits from vastly more pretraining data and a broader semantic curriculum, yielding representations better aligned for cross-document similarity. In contrast, our domain model is smaller, trained only with MLM, and lacks contrastive or retrieval-aware supervision. Sliding windows (512 with overlap) increase exposure to long texts but add little new signal, as overlapping segments are semantically redundant and averaging them can dilute discriminative cues. Empirically, the general embedder retrieves neighbors with stronger diagnostic tag overlap even under strict thresholds, showing that scale and general semantics dominate domain-only MLMs on this task.

Absolute nDCG values represent conservative lower bounds rather than poor utility, as keyword overlap is a noisy and incomplete gold standard. Despite this, head metrics (P@k, HitRate@k, RBP) and bootstrap tests show large, statistically significant robust improvements for the OpenAI embedder across all cutoffs and thresholds. These results indicate that pretraining scale and semantic breadth outweigh domain restriction under noisy labels. Future work will explore paragraph-level and hybrid sparse–dense retrieval, retrieval-aware pretraining, and expert-based evaluation to better capture legal relevance.

## 6. Acknowledgement

Tereza Novotná acknowledges the support of National Recovery Plan funded project MPO 60273/24/21300/21000 CEDMO 2.0 NPO. Jakub Harašta acknowledges the support of the Grant Agency of Masaryk University (GAMU) project "Forensic Support for Building Trust in Smart Software Ecosystems" (no. MUNI/G/1142/2022).

## References

[1] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781. 2013.

[2] Le Q, Mikolov T. Distributed representations of sentences and documents. In: International conference on machine learning. PMLR; 2014. p. 1188-96.
[3] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers); 2019. p. 4171-86.
[4] Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I. LEGAL-BERT: The muppets straight out of law school. arXiv preprint arXiv:201002559. 2020.
[5] Mandal A, Ghosh K, Ghosh S, Mandal S. Unsupervised approaches for measuring textual similarity between legal court case reports. Artificial Intelligence and Law. 2021;29(3):417-51.
[6] Westermann H, Savelka J, Benyekhlef K. Paragraph similarity scoring and fine-tuned BERT for legal information retrieval and entailment. In: JSAI International Symposium on Artificial Intelligence. Springer; 2020. p. 269-85.
[7] Vuong YTH, Bui QM, Nguyen HT, Nguyen TTT, Tran V, Phan XH, et al. SM-BERT-CR: a deep learning approach for case law retrieval with supporting model. Artificial Intelligence and Law. 2023;31(3):601-28.
[8] Mentzingen H, António N, Bacao F. Effectiveness in retrieving legal precedents: exploring text summarization and cutting-edge language models toward a cost-efficient approach. Artificial Intelligence and Law. 2025:1-21.
[9] Tang Y, Qiu R, Huang Z. UQLegalAI@ COLIEE2025: Advancing Legal Case Retrieval with Large Language Models and Graph Neural Networks. arXiv preprint arXiv:250520743. 2025.
[10] Savelka J, Ashley KD. The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts. Frontiers in Artificial Intelligence. 2023;6:1279794.
[11] Lee J, Lee G. Long context window-based zero-shot legal interpretation of building codes and regulations. Automation in Construction. 2025;179:106450.
[12] Dugac G, Altwicker T. Classifying legal interpretations using large language models. Artificial Intelligence and Law. 2025:1-19.
[13] Guha N, Nyarko J, Ho D, Ré C, Chilton A, Chohlas-Wood A, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. Advances in neural information processing systems. 2023;36:44123-279.
[14] Guo X, Huang Y, Wei B, Kuang K, Wu Y, Gan L, et al. Specialized or general AI? a comparative evaluation of LLMs' performance in legal tasks. Artificial Intelligence and Law. 2025:1-37.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251606*

# PONK: Tool for Client-Oriented Legal Writing in Czech

Tereza NOVOTNÁ [a,1] , Jan ČERNÝ [a] , Ivan KRAUS [a] , Ivana KVAPILÍKOVÁ [a] , Jiří MÍROVSKÝ [a] , Arnold STANOVSKÝ [a] and Barbora HLADKÁ [a]

[a] *Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague, CZ*

**Abstract.** Administrative and legal communication is often difficult for laypersons to understand, creating barriers to justice and undermining trust in public institutions. The **PONK** tool helps authors identify and revise unclear text using a multimodular approach that combines *linguistic rules* and *lexical surprisal* based on large language models. This integration ensures precise and adaptable detection of problematic passages. Human evaluation confirmed the tool's effectiveness in improving legal text comprehensibility. PONK's user-friendly interface highlights unclear segments at the word level, offering a practical solution for clearer and more accessible legal communication.

**Keywords.** readability, legal texts, lexical surprisal, linguistic analysis, large language models

## 1. Introduction

Improving the readability of legal and administrative language is vital for access to justice, as unclear communication can hinder citizens from understanding their rights and erode trust in institutions. This study introduces PONK—an open, modular web application that combines linguistic and machine-learning methods to evaluate and enhance the clarity of legal texts. PONK assesses how well citizens can understand common administrative documents such as fines or benefit applications. The Czech Ombudsman's 2023 Annual Report recorded about 7 000 complaints concerning unclear official communication,[2] exceeding available assistance. Clearer texts could therefore reduce complaints and strengthen public trust.

The tool highlights linguistically complex segments and suggests reformulations. Its backend integrates two modules—*lexical surprisal*, *linguistic rules*—combining linguistic analysis with LLM-based semantic assessment. The modular design enables flexible, fine-grained feedback at the sub-sentence level, offering actionable guidance rather than generic readability scores. Figure 1 shows the system overview. The paper proceeds with related work (Section 2), data (Section 3), system design (Section 4), module evaluation (Section 5) and conclusions (Section 6).

[1] Corresponding author. Email: novotna@ufal.mff.cuni.cz.

[2] Available at: https://www.ochrance.cz/dokument/zpravy_pro_poslaneckou_snemovnu_2023/vyrocni-zprava-2023.pdf

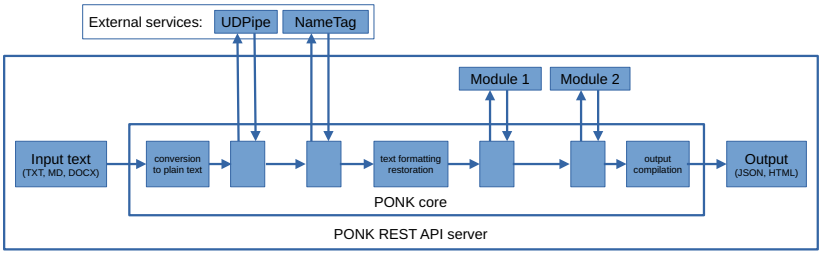



**Figure 1.** PONK Application Architecture

## 2. Related Work

Readability reflects how easily a text can be processed, while understandability also requires correct interpretation of meaning [1]. Many plain-language metrics applied to legal texts overlook these deeper aspects. Automatic formulas based on syntax or vocabulary are scalable but correlate weakly with human judgment [2,3], whereas human methods (cloze tests, crowdsourcing) better capture comprehension at higher cost [2,4].

Legal readability research, rooted in the plain-language movement [5], extends to contracts [6], legislation [7], and legal websites [8]. PONK combines general-language readability models with domain-specific linguistic expertise (Section 4). It integrates two complementary approaches: *lexical surprisal*, which links unexpected words to cognitive load and reading difficulty [9,10,11], and *linguistic rules*, which evaluate structural and lexical features such as sentence length, dependency depth, coherence markers, and word frequency [12,13,14]. Together, these methods capture both probabilistic and rule-based sources of legal text complexity.

## 3. Data

The **KUK 0.0 corpus**[3] was developed specifically for PONK and serves as the foundation for both modules. It is a pilot collection of Czech legal and administrative texts designed for the manual and automatic assessment of readability and understandability, with evaluation to follow in later stages.

KUK 0.0 comprises three subcorpora: (1) *Public materials of Frank Bold*, a nonprofit organization offering plain-language legal guidance for citizens; (2) *Statements of the Ombudsman*, opinions addressing specific complaints from the public (e.g., in asylum or social security matters); and (3) *Information flyers of the Ombudsman*, brief educational texts explaining recurring legal issues.

Together, these subcorpora include 6,051 documents and 19.2 million words. KUK 0.0 also incorporates the Czech Court Decisions Corpus (CzCDC 1.0) [15], containing rulings of the Supreme, Supreme Administrative, and Constitutional Courts. The KUK 0.0 corpus is unique in classifying Czech legal documents by their degree of understandability, enabling both PONK development and its future evaluation on expert and citizen-oriented texts.

[3] https://lindat.mff.cuni.cz/repository/items/29452e2c-33eb-47a0-8641-f9c82088adaa

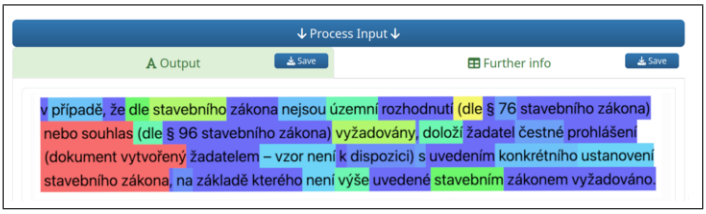



**Figure 2.** Visual representation of lexical surprisal

## 4. System Architecture

PONK is a modular client–server application with a web interface and REST API backend. The Perl-based core coordinates internal modules and external NLP tools via REST using the CoNLL-U format (see Figure 1). External services include UDPipe [16] for tokenization, lemmatization, and syntactic parsing, and NameTag [17] for named entity recognition. Internally, two modules operate in sequence: the *Lexical Surprisal Module*, which estimates word-level predictability, and the *Linguistic Rules Module*, which applies syntax-based readability checks.

The web client allows users to paste or upload text, submit it via REST, and interactively explore results. PONK can also be run as a command-line utility or accessed through a public REST API.[4] The public instance, available for free non-commercial use, ensures privacy by not storing processed texts; only minimal metadata (timestamp, data size, configuration, and IP address) is logged, and all external services maintain the same anonymity policy.

### *4.1. Lexical Surprisal Module*

Lexical surprisal quantifies how unexpected a word is in context, reflecting the cognitive effort required for comprehension [18,19]. High surprisal indicates low readability. We adapted the GLTR visualization tool [20]—originally designed to detect AI-generated text—into *Glitter*, which highlights word-level surprisal as an indicator of textual difficulty.

Input text is tokenized using `Czech GPT-2 XL`[5], whose 64k-token vocabulary splits most words into one to three subword units. The model predicts each token from up to 128 preceding tokens and computes its probability, producing surprisal scores for every word. These are color-coded by probability bands, with alternative likely continuations shown. The visualization helps legal authors spot cognitively demanding segments and revise them for better readability.

### *4.2. Linguistic Rules Module*

We implemented 35 linguistic rules inspired by Czech plain-language handbooks [21, 22,23,24]. These works emphasize clarity, simplicity, and avoidance of ambiguous or cognitively demanding structures. Several rules address the so-called *garden-path ef-*

[4] https://quest.ms.mff.cuni.cz/ponk/
[5] https://huggingface.co/BUT-FIT/Czech-GPT-2-XL-133k

*fect*—temporary misinterpretations that arise when a grammatical sentence allows multiple parses until disambiguated [25].

All rules are implemented in Python using Udapi[6] blocks, relying on UDPipe and NameTag annotations and, where needed, MorphoDiTa [26] for morphological paradigms. When a parsed subtree matches a rule pattern, relevant tokens are labeled with the rule name, role-specific *participants*, and a unique ID identifying the instance. Rules are grouped into five categories for interpretability: **ambiguity** (10 rules, e.g., garden-path constructions), **clusters** (5 rules, frequent patterns lowering readability, e.g., excess negations), **phrases** (8 rules, e.g., weak-meaning words), **structural** (8 rules, e.g., distant predicates), and **acceptability** (4 rules, e.g., incomplete conjunctions).

In the interface, each category is visualized with distinct highlight colors, with grey marking overlaps. Current limitations include the inability to process larger textual units beyond sentences and the lack of deeper lexical-semantic understanding—factors that restrict the precision of some cluster and semantic rules.

## 5. Evaluation

### *5.1. Evaluation of Lexical Surprisal Module*

We qualitatively evaluated the lexical surprisal module on 50 randomly selected documents from the KUK 0.0 corpus, using its visual interface where word surprisal is color-coded from blue (low) to red (high). The analysis, conducted by legal and linguistic experts in our team, focused on how these gradients correspond to perceived readability (see Figure 2). Although exploratory, the evaluation provided useful insights into interpretability and model behavior.

Low-surprisal words (dark blue) often appear in fixed expressions, indicating fluency but sometimes redundancy, while high-surprisal words (red) frequently disrupt sentence flow. Surprisal naturally decreases as the context lengthens. Similar to [20], we observed that a balanced mix of predictable and unexpected words produces natural, comprehensible text, whereas excessive unpredictability reduces clarity. Contrary to expectations, specialized legal terms often received low surprisal—language models predict them easily due to their presence in large training corpora. Future work will adapt prompting strategies to better reflect lay comprehensibility while retaining contextual accuracy.

### *5.2. Evaluation of Linguistic Rules Module*

We evaluated the linguistic rules on 30 texts from the KUK 0.0 corpus, including Ombudsman materials and court decisions from the ConCo and SupAdmCo subcorpora of CzCDC 1.0 (see Section 3). The rules were applied and visually inspected using the Brat Rapid Annotation Tool.[7] A legally trained annotator assessed each rule occurrence as *useful* (improves readability), *not_useful* (captures the phenomenon but not helpful), or *incorrect* (false positive).

Of the 35 implemented rules, 33 appeared in the dataset, yielding 5,069 annotations (10–735 per rule). Two rules from the *acceptability* category remain untested. Most anno-

[6] https://udapi.github.io/
[7] https://brat.nlplab.org/

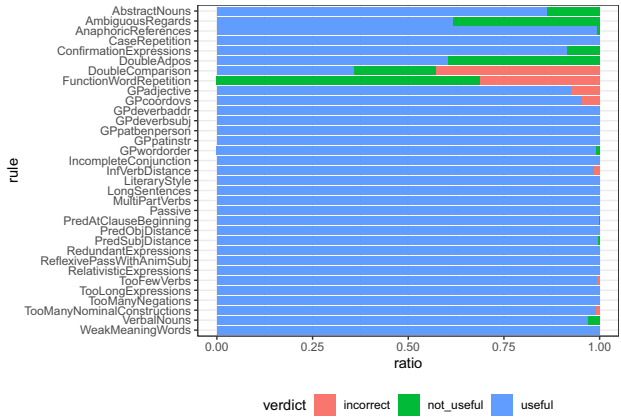



**Figure 3.** Evaluation of Linguistic Rules Module (ratios of annotation outcomes)

tations were judged *useful*, confirming that the rules identify relevant readability issues, though some strict conditions may reduce recall. Cases labeled *not_useful* often involved named entities, idioms, or canonical legal expressions that are stylistically fixed rather than unclear. These findings support the overall validity of the module while highlighting the need for semantic filtering to reduce over-triggering.

## 6. Conclusion and Future Work

PONK integrates two complementary methods—lexical surprisal and linguistically informed rules—enhanced by expert knowledge of Czech legal drafting (see Section 4). These modules provide real-time feedback at the word level, effectively identifying unclear passages. Qualitative evaluation confirmed their usefulness, though broader testing on larger datasets is needed.

Future work will compare surprisal patterns in original and simplified legal texts, refine rule thresholds through genre-specific profiles, and explore multiple language models for surprisal estimation. A planned third, supervised module will assess textual coherence and relevance, extending PONK toward deeper semantic evaluation. Its modular architecture enables seamless integration of such advances, ensuring ongoing alignment with research in legal readability.

## Acknowledgement

The research has been supported by the Technology Agency of the Czech Republic under the project TQ01000526 (9/2023-12/2025).

## References

[1] Curtotti M, McCreath E. A right to access implies a right to know: An open online platform for research on the readability of law. J Open Access L. 2013;1:1.

[2] De Clercq O, Hoste V, Desmet B, Van Oosten P, De Cock M, Macken L. Using the crowd for readability prediction. Natural Language Engineering. 2014;20(3):293-325.
[3] Heydari P, Riazi AM. Readability of texts: Human evaluation versus computer index. Mediterranean Journal of Social Sciences. 2012;3(1):177-90.
[4] Munro R, Bethard S, Kuperman V, Lai VT, Melnick R, Potts C, et al. Crowdsourcing and language studies: the new generation of linguistic data. In: NAACL Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk. Association for Computational Linguistics; 2010. p. 122-30.
[5] Benoliel U, Becher SI. The duty to read the unreadable. BCL Rev. 2019;60:2255.
[6] Arbel YA. The readability of contracts: Big data analysis. Journal of Empirical Legal Studies. 2024;21(4):927-78.
[7] Curtotti M, McCreath E, Bruce T, Frug S, Weibel W, Ceynowa N. Machine learning for readability of legislative sentences. In: Proceedings of the 15th International Conference on Artificial Intelligence and Law; 2015. p. 53-62.
[8] Dyson DD, Schellenberg K. Access to justice: The readability of legal services corporation legal aid internet services. Journal of poverty. 2017;21(2):142-65.
[9] van Schijndel M, Linzen T. Can entropy explain successor surprisal effects in reading? arXiv preprint arXiv:181011481. 2018.
[10] Monsalve IF, Frank SL, Vigliocco G. Lexical surprisal as a general predictor of reading time. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics; 2012. p. 398-408.
[11] Staub A. Predictability in language comprehension: Prospects and problems for surprisal. Annual Review of Linguistics. 2024;11.
[12] Kleijn S. Clozing in on readability: How linguistic features affect and predict text comprehension and on-line processing. LOT; 2018.
[13] Hackemann T, Heine L, Höttecke D. Challenging to read, easy to comprehend? Effects of linguistic demands on secondary students' text comprehension in physics. International Journal of Science and Mathematics Education. 2022;20(Suppl 1):43-68.
[14] Frantz RS, Starr LE, Bailey AL. Syntactic complexity as an aspect of text complexity. Educational Researcher. 2015;44(7):387-93.
[15] Novotná T, Harašta J. Czech Court Decisions Corpus (CzCDC 1.0); 2019. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. Available from: http://hdl.handle.net/11372/LRT-3052.
[16] Straka M. UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Brussels, Belgium: Association for Computational Linguistics; 2018. p. 197-207. Available from: https://www.aclweb.org/anthology/K18-2020.
[17] Straková J, Straka M, Hajic J. Neural Architectures for Nested NER through Linearization. In: Korhonen A, Traum D, Màrquez L, editors. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics; 2019. p. 5326-31. Available from: https://aclanthology.org/P19-1527.
[18] Fernandez Monsalve I, Frank SL, Vigliocco G. Lexical surprisal as a general predictor of reading time. In: Daelemans W, editor. Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon, France: Association for Computational Linguistics; 2012. p. 398-408. Available from: https://aclanthology.org/E12-1041/.
[19] Cinková S. Linguistic Factors in the Readability of Czech Administrative and Legal Texts. Vienna, Austria: Praesens Verlag; 2024.
[20] Gehrmann S, Strobelt H, Rush AM. GLTR: Statistical Detection and Visualization of Generated Text. CoRR. 2019;abs/1906.04043. Available from: http://arxiv.org/abs/1906.04043.
[21] Šamánková J, Kubíková B, editors. Jak psát srozumitelné úřední texty. Příručka srozumitelného psaní pro úředníky. Brno: Kancelář veřejného ochránce práv; 2022. Available from: https://www.ochrance.cz/uploads-import/ESO/pruka/Prirucka_srozumitelneho_psani_tisk.pdf.
[22] Šváb J. Jak psát, aby se to dalo číst. Příručka přístupného psaní. Prague: Leges; 2021.
[23] Sgall P, Panevová J. Jak psát a jak nepsát česky. Prague: Karolinum; 2014.
[24] Chromý J, Cehákova M. Diversity of garden-path structures - SPR. OSF Registries; 2023.
[25] Trueswell JC, Tanenhaus MK, Kello C. Verb-Specific Constraints in Sentence Processing: Separating Effects of Lexical Preference From Garden-Paths. Journal of Experimental Psychology: Learning,

Memory, and Cognition. 1993;19(3):528-53.

[26] Straková J, Straka M, Hajič J. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Baltimore, Maryland: Association for Computational Linguistics; 2014. p. 13-8. Available from: http://www.aclweb.org/anthology/P/P14/P14-5003.pdf.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*

*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251607*

# EmoLawConnect: An AI-Powered Matching Engine for Professionals Supporting Family Conflict Resolution

Kohei OSHIO [a,1]
[a] *Graduate School of Information and Communication, Meiji University, Japan*

**Abstract.** *EmoLawConnect* is an AI-powered matching engine for family conflict resolution that recommends lawyers by combining legal case profiles with clients' emotional preferences and value alignment. It integrates a multi-layer pipeline that processes user inputs (e.g., emotional expression, legal needs) and professional profiles (e.g., legal expertise, practice orientation). The architecture merges rule-based pre-filtering with a semantic–emotional scoring module built on matrix factorization and value-aware alignment. The prototype was evaluated through scenario-based testing and a rule-based legal-consistency proxy. Results suggest improved user–professional compatibility and trust, with potential contributions to access to justice. The approach provides a foundation for broader legal-service applications where affective or interpersonal alignment matters.

**Keywords.** AI and law, legal tech, professional matching, access to justice, family dispute resolution, value alignment

## 1. Introduction

Access to legal professionals in family-related conflicts (e.g., divorce, custody, inheritance) requires not only legal expertise but also sensitivity to emotional and relational dynamics. Existing legal tech often focuses on document automation or rule-based guidance, yet rarely considers interpersonal compatibility between users and professionals, which can materially affect negotiation outcomes.

Building on our prior work on a negotiation-support chatbot that estimates both Legal-Emotional BATNA [1], and on follow-up interviews with former judges and practicing attorneys in Japanese family law, we emphasize the need for detailed user profiling. Relevant attributes include family structure, marital history, conflict patterns, personality traits, and communication styles [2,3].

To comply with Article 72 of Japan's Attorney Act, *EmoLawConnect* (i) does not provide legal advice, (ii) does not broker legal services for compensation, and (iii) presents outputs as internal decision support for licensed professionals. UI flows and contractual terms ensure that final client-engagement decisions remain solely with attorneys and law firms.

[1] Email (until Mar 2026): koshio@meiji.ac.jp; Permanent: oshio.k.contact22@gmail.com

This paper details the system architecture, matching algorithm design, and a preliminary evaluation framework for *EmoLawConnect*. The current prototype focuses on rule-based matching using structured user and expert profiles. Future work includes human-in-the-loop validation with domain experts, calibration analysis, and evaluation on anonymized real-world datasets within existing referral workflows.

**Contributions.** (i) a rule-isolated proxy for legal consistency, (ii) an affect/value diagnostic kept out of supervision, and (iii) a compact neural scorer that outperforms MF on the proxy task.

## 2. Related Work

Recommender systems infer user preferences from behavioral traces across domains such as e-commerce, media, and healthcare. Legal settings pose distinct challenges: matters are sensitive, intents vary widely, and professional profiles are heterogeneous. Early legal recommenders relied on ontology-based matching between user problems and legal knowledge, prioritizing explainability over personalization [4].

Data-driven approaches later modeled implicit compatibility via collaborative filtering (CF); matrix factorization (MF) projects users and professionals into a shared latent space to capture affinities [7]. Incorporating side information, profiles, and context improves over pure CF [6]. Surveys codify evaluation protocols and ranking metrics [14,18], while MNAR biases warn against naïve train/test designs [17].

In emotionally charged domains (e.g., family law), behavioral traces alone are insufficient. Affective/personality-aware recommenders suggest that value/affect congruence boosts satisfaction and trust [16,19]. Legal-tech for vulnerable users highlights the need for empathic, careful UX [5].

Beyond accuracy, high-stakes contexts require attention to ethics and process. Fairness-aware recommendation studies exposure/representation via constraints and reranking [9,10,11]; work on interpretability and legitimacy cautions against naïve transparency [12,8,13].

We also connect to legal-aid triage and referral systems reporting operational constraints in intake, eligibility, and routing; our intake minimalism borrows from these deployments while adding an affect-aware diagnostic layer [20,21,22,23,24].

*EmoLawConnect* contributes at this intersection by: (i) combining legal specialization with affective–value signals; (ii) treating the Emotional-Aware Alignment Score (EAS) as a *diagnostic* (not supervisory) signal within the evaluation pipeline; and (iii) aligning optimization with legal practice by evaluating against a rule-based legal-consistency proxy instead of an expert gold set, thereby isolating legal constraints from affective alignment in the current prototype.

## 3. System Architecture

Our system has three parts: user profiling, professional representation, and a scorer that predicts user–professional compatibility.

As a baseline, we use biased matrix factorization (MF). Let $u \in \{1,\dots,M\}$ and $i \in \{1,\dots,N\}$ index users and professionals, and let $\mathscr{K}$ be the set of observed pairs. Let $\mu$ be the global mean, $b_u, b_i$ the biases, and $\mathbf{p}_u, \mathbf{q}_i \in \mathbb{R}^k$ the latent factors. The estimate is

$$\hat{R}_{ui} = \mu + b_u + b_i + \mathbf{p}_u^\top \mathbf{q}_i. \tag{1}$$

Parameters are learned by minimizing squared error over $\mathcal{K}$ with $\ell_2$ regularization on user and professional terms:

$$\min_{b,\mathbf{p},\mathbf{q}} \sum_{(u,i)\in\mathcal{K}} (R_{ui} - \hat{R}_{ui})^2 + \lambda \left( \sum_u (\|\mathbf{p}_u\|^2 + b_u^2) + \sum_i (\|\mathbf{q}_i\|^2 + b_i^2) \right). \tag{2}$$

Affective and value-related compatibility is quantified with a cosine score. Let $v_u, p_i \in \mathbb{R}^m$ be the user's value–affective vector and the professional's psychological/value vector, both optionally $\ell_2$-normalized:

$$\mathrm{EAS}(u,i) = \frac{\langle v_u, p_i \rangle}{\|v_u\| \, \|p_i\|} \in [-1,1]. \tag{3}$$

EAS is excluded from supervision and used only in diagnostics/ablations to avoid label–feature coupling.

The main scorer is a multilayer perceptron that consumes structural and affect-aware inputs. Let $e_u$ be the user's legal-need embedding, $e_i$ the professional's profile embedding, and $a_u$ the user's affect features (valence = settlement preference; arousal = pace tolerance). Let $\Theta$ denote all trainable parameters, and let $\lambda > 0$ be the $\ell_2$ regularization weight. The match probability is

$$\hat{y}_{ui} = \mathrm{MLP}\big([\, e_u \,\|\, e_i \,\|\, a_u \,]\big), \tag{4}$$

and training minimizes the average binary cross-entropy with $\ell_2$ regularization:

$$\mathscr{L} = -\frac{1}{|\mathcal{K}|} \sum_{(u,i)\in\mathcal{K}} \Big[ y_{ui} \log \hat{y}_{ui} + (1 - y_{ui}) \log(1 - \hat{y}_{ui}) \Big] + \lambda \|\Theta\|_2^2. \tag{5}$$

When explicitly noted in ablations, we append $\mathrm{EAS}(u,i)$ to the input in Eq. 4; otherwise the scorer runs without EAS.

We later combine the scorer with a lightweight justification module that generates short natural-language rationales from dialogue metadata and injects their embeddings back into the scorer as an auxiliary input to the ranker (see Discussion).

## 4. Evaluation Design

We evaluate *EmoLawConnect* along three questions: accuracy on ranking/classification metrics, consistency with a rule-based legal proxy, and the diagnostic relation between affect/value alignment (EAS) and successful matches. We report Precision@K, Recall@K, F1, Hit Rate, NDCG, and ROC AUC [14,17,18]. Legal appropriateness is measured with a rule-based proxy $L(u,i) \in \{0,1\}$ (specialization, jurisdiction/availability, conflict-of-interest); no human labels are used. We emphasize threshold-agnostic summaries (AUC, NDCG@K). EAS is excluded from supervision and used only post hoc for diagnostics/ablations to avoid label–feature coupling.

The dataset comprises 50 users and 30 professionals. Each user has a legal-need embedding, a value vector, and affect features; each professional has a profile embedding and a psychological/value vector. We use five user-stratified folds (seed 42). Any operating thresholds used in diagnostics (e.g., Top-$K$ summaries) are selected on training folds only. Affect features $a_u$ are elicited via a short survey (e.g., 5-point valence and arousal items, and preferred communication style), normalized to $[0,1]$ and concatenated with the legal-need embedding. This keeps EAS diagnostic while ensuring that supervised signals derive only from the legal proxy.

The scorer is an MLP over $[e_u \| e_i \| a_u]$ (EAS disabled in the main results; a (+EAS) variant appears only in ablations), trained with the average binary cross-entropy and $\ell_2$ regularization in Eq. 5. Unless noted, the defaults are embedding size 64, Adam with learning rate $10^{-3}$, $\lambda = 0.01$, batch size 32, up to 100 epochs, and early stopping with patience 5. We report means over five folds.

## 5. Implementation and Evaluation Results

We evaluate on the synthetic dataset above using only the rule-based labels $L(u,i)$ for supervision. The baseline is biased MF; the proposed model is the MLP in Eq. 4 trained with Eq. 5. Unless explicitly stated in ablations, EAS is disabled as an input feature.

**Table 1.** Recommendation accuracy against the rule-based proxy $L(u,i)$ (mean over five user-stratified folds; seed=42).

| Metric | Baseline (MF) | Proposed (MLP) |
|---|---|---|
| Precision@5 | 0.44 | 0.68 |
| Recall@5 | 0.30 | 0.74 |
| F1-score@5 | 0.35 | 0.71 |
| Hit Rate@5 | 0.52 | 0.84 |
| NDCG@5 | 0.50 | 0.79 |
| ROC AUC | 0.74 | 0.88 |

The MLP outperforms MF across all metrics at $K = 5$. Agreement with the legal proxy is summarized by ROC AUC (Table 1); Top-$K$ or fixed thresholds are shown only as illustrative operating points. As a post-hoc diagnostic, EAS (Eq. 3) exhibits a higher mean for correctly matched pairs, suggesting that affective–value congruence correlates with success; these analyses do not affect training targets. Probability calibration is not evaluated in this version.

## 6. Discussion

The results suggest that *EmoLawConnect* can rank suitable professionals well under a rule-based legal proxy, while EAS provides an interpretable signal of affective fit. Gains on Precision@5, NDCG, and ROC AUC indicate that combining structured profiles with a lightweight neural scorer is effective; EAS is informative diagnostically but is kept outside supervision to avoid leakage.

Limitations remain. The synthetic dataset cannot capture the full social and interpersonal context of family law. Fixed-length affect/value vectors simplify dynamic states.

Future evaluations should combine expert review, small-signal field studies within referral workflows, and calibration analysis.

Architecture-wise, the system is modular. A justification module can generate short natural-language rationales from dialogue and metadata, and these rationales can be embedded and fed back to the scorer as an auxiliary feature. The same modularity extends to many-to-many settings (e.g., lawyer–counselor–mediator teams) by scoring candidate teams with both legal constraints and average EAS, then optimizing under workload and diversity constraints.

Ethically, deployment must prioritize procedural legitimacy: fairness-aware exposure, transparent operating points, and user-facing explanations proportionate to risk [8,9,10,11]. These concerns are integral to design, not an afterthought.

Data collection in referral workflows is constrained: users and professionals provide sparse, strategic, and sometimes inconsistent inputs. We therefore scope the supervised target to a legal proxy and treat affective alignment as diagnostic, while designing the intake to minimize burden ($\leq$10 items, defaultable answers) and mirror existing referral forms.

As a next step, we will run a pilot ($\approx$15–20 cases) inside an existing referral workflow, collecting user-reported satisfaction, perceived fit, 7-day intent-to-retain, expert agreement (Cohen's $\kappa$), and proxy agreement (AUC/NDCG).

## 7. Conclusion and Future Work

We presented *EmoLawConnect*, a matching engine for family-law services that pairs legal relevance with affect/value considerations. Under a rule-based legal proxy, a simple MLP over structured profiles outperformed an MF baseline, and EAS offered a useful diagnostic of interpersonal fit.

Next steps focus on realism and governance: (i) expert-reviewed and user-involved evaluations with anonymized real cases, (ii) calibration and robustness checks, (iii) multimodal profiling that augments surveys with text-derived embeddings, and (iv) team matching with fairness and workload constraints. The goal is not to automate legal judgment, but to improve human–human matching in emotionally complex settings while respecting legal and ethical boundaries.

## Acknowledgements

This research was supported by The Foundation for Research on Civil Dispute Resolution under the following funded projects: (1) "Creative Negotiation Support in ODR Using Spiking Neural Networks (SNN)," and (2) "An Innovative Approach to Mediation Chatbots Utilizing Mathematical Science-Based Methods."

## References

[1] Oshio K. Legal-Emotional BATNA: AI chatbot addressing divorce legalities and emotional complexities, and research of social implementation in Japan. In: Proceedings of JURIX 2024: The Thirty-seventh Annual Conference. IOS Press; 2024. p. 333–337. https://doi.org/10.3233/FAIA241264

[2] Oshio K. Legal-Emotional BATNA: AI chatbot addressing divorce legalities and emotional complexities, and research of social implementation in Japan. Meiji University Graduate School Journal of Information and Communication. 2025;25:1–21. https://meiji.repo.nii.ac.jp/records/2001951

[3] Oshio K. Towards human-centered legal AI: Designing a facilitative mediation agent integrating tacit knowledge and multimodal dialogue cues. In: ICAIL 2025 Doctoral Consortium. Northwestern University; 2025. p. 1–10. https://bpb-us-e1.wpmucdn.com/sites.northwestern.edu/dist/c/8101/files/2025/06/ICAIL25_DC_Kohei-Oshio-_final-1.pdf

[4] Breuker J, Elhag A, Petkov E, Winkels R. Ontologies for legal information serving and knowledge management. In: Proceedings of JURIX 2002. IOS Press; 2002. p. 73–82. https://jurix.nl/pdf/j02-08.pdf

[5] Awasekar DD, Lobo LMRJ. Empowering women through AI: A comprehensive chatbot for domestic violence awareness and legal support in India. In: Auer ME, Rüütmann T, editors. *Futureproofing Engineering Education for Global Responsibility* (ICL 2024). Lecture Notes in Networks and Systems, Vol. 1281. Springer; 2025. p. 198–205. https://doi.org/10.1007/978-3-031-83520-9_19

[6] Yu E, Wang C, Zou X, You C, He X. Lawyer recommendation system based on user profiles and collaborative filtering. Advances in Engineering Technology Research. 2024;11(1):237–244. https://doi.org/10.56028/aetr.11.1.237.2024

[7] Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems. Computer. 2009;42(8):30–37. https://doi.org/10.1109/MC.2009.263

[8] Binns R, Veale M, Van Kleek M, Shadbolt N. "It's reducing a human being to a percentage": Perceptions of justice in algorithmic decisions. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM; 2018. Article 377. https://doi.org/10.1145/3173574.3173951

[9] Burke R. Multisided fairness for recommendation. In: Proceedings of the Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2017); 2017.

[10] Mehrotra R, McInerney J, Bouchard H, Lalmas M, Diaz F. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness and satisfaction in recommendation systems. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18). ACM; 2018. p. 2243–2252. https://doi.org/10.1145/3269206.3272027

[11] Sonboli N, Burke R, Mobasher B. Fairness-aware recommendation with LibRec-Auto. In: Proceedings of the 14th ACM Conference on Recommender Systems (RecSys '20). ACM; 2020. p. 594–596. https://doi.org/10.1145/3383313.3418481

[12] Cowgill BO, Dell'Acqua F, Deng S, Hsu D, Verma N, Chaintreau A. Biased programmers? Or biased data? A field experiment in operationalizing AI ethics. In: Proceedings of the 21st ACM Conference on Economics and Computation (EC '20). ACM; 2020. p. 679–681. https://doi.org/10.1145/3391403.3399545

[13] Ananny M, Crawford K. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society. 2018;20(3):973–989. https://doi.org/10.1177/1461444816676642

[14] Gunawardana A, Shani G. Evaluating recommender systems. In: Ricci F, Rokach L, Shapira B, Kantor PB, editors. *Recommender Systems Handbook*. 2nd ed. Springer; 2015. p. 265–308. https://doi.org/10.1007/978-1-4899-7637-6_8

[15] Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960;20(1):37–46. https://doi.org/10.1177/001316446002000104

[16] Hu R, Pu P. Enhancing collaborative filtering systems with personality information. In: Proceedings of the 5th ACM Conference on Recommender Systems (RecSys '11). ACM; 2011. p. 197–204. https://doi.org/10.1145/2043932.2043967

[17] Steck H. Training and testing of recommender systems on data missing not at random. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '10). ACM; 2010. p. 713–722. https://doi.org/10.1145/1835804.1835895

[18] Zangerle E, Bauer C. Evaluating recommender systems: Survey and framework. ACM Computing Surveys. 2022;55(8):170:1–170:38. https://doi.org/10.1145/3556536

[19] Katarya R, Verma OP. Recent developments in affective recommender systems. Physica A: Statistical Mechanics and its Applications. 2016;461:182–190. https://doi.org/10.1016/j.physa.2016.05.046

[20] State Court Administrative Office. Independent evaluation of Michigan's intake and referral system. Michigan Judiciary; 2023. https://www.courts.michigan.gov/

494822/siteassets/committees%2C-boards-special-initiatves/justiceforall/ecosytemevaluationreport-final.pdf

[21] Coumarelos C, McDonald HM. Developing a triage framework: linking clients with services at Legal Aid NSW. Law and Justice Foundation of NSW; 2019. https://lawfoundation.net.au/developing-a-triage-framework-linking-clients-with-services-at-legal-aid-nsw/

[22] Sandefur RL, Smyth AC, Anderson T, et al. Roles Beyond Lawyers: Evaluation of the New York City Court Navigators Program. American Bar Foundation; 2016. https://www.americanbarfoundation.org/wp-content/uploads/2024/03/new_york_city_court_navigators_report_final_with_final_links_december_2016.pdf

[23] Greiner DJ, Pattanayak CW. Randomized evaluation in legal assistance. Yale Law Journal. 2012. https://www.yalelawjournal.org/pdf/1088_q6v2s7cc.pdf

[24] Zorza R. The Access to Justice "Sorting Hat": Towards a System of Triage for Legal Services. Denver University Law Review. 2012;89(4). https://digitalcommons.du.edu/dlr/vol89/iss4/6/

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251608*

# Defending the Hierarchical Result Models of Precedential Constraint

Henry PRAKKEN [a,1] and Wijnand VAN WOERKOM [b]
[a] *Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands*
[b] *Max Planck Institute for Comparative and International Private Law, Hamburg, Germany*

**Abstract.** In recent years, hierarchical case-based-reasoning models of precedential constraint have been proposed. Trevor Bench-Capon criticised these models, among other things, on the grounds that they would not account for the possibility that intermediate factors are established with different strengths by different base-level factors. In this paper we respond to these criticisms for van Woerkom's result-based hierarchical models. We argue that in some examples Bench-Capon seems to interpret intermediate factors as dimensions, and that applying van Woerkom's dimension-based version of the hierarchical result model to these examples avoids Bench-Capon's criticisms.

**Keywords.** case-based reasoning, precedential constraint, intermediate factors

## 1. Introduction

In [7] a hierarchical factor-based model of precedential constraint was proposed, extending the 'flat' factor-based result and reason models of precedential constraint of [5]. Very briefly, the flat model represents *cases* as a binary decision plus a *fact situation*, which consists of two sets of factors pro and con the decision. A decision in a new fact situation is *forced* if a precedent for that decision exists that cannot be distinguished from the new fact situation, i.e., if the latter contains at least all pro-decision factors of the precedent and at most all con-decision factors of the precedent. In [7] this is refined in terms of a factor hierarchy [1], in which *basic* factors are pro or con more *abstract* factors, which ultimately are pro or con the final binary decision. This allows to make the forcedness of the final decision relative to intermediate decisions instead of to basic factors.

In [2,3], Trevor Bench-Capon criticised the resulting model, among other things, on the ground that it does not account for the possibility that intermediate factors are established with different strengths by different base-level factors. In this paper we argue that in some examples Bench-Capon seems to interpret intermediate factors as dimensions, and that applying [6]'s dimension-based version of the hierarchical result model to these examples avoids Bench-Capon's criticisms.

[1] Corresponding Author: Henry Prakken, h.prakken@uu.nl.

## 2. Examples and Bench-Capon's criticisms

Bench-Capon discusses two examples relevant for us about a family with parents Jack and Jo with two children Emma and Max, where the parents want to consistently decide about whether their children can have ice cream. Bench-Capon assumes the factor hierarchy depicted in Figure 1, adapted from [4].

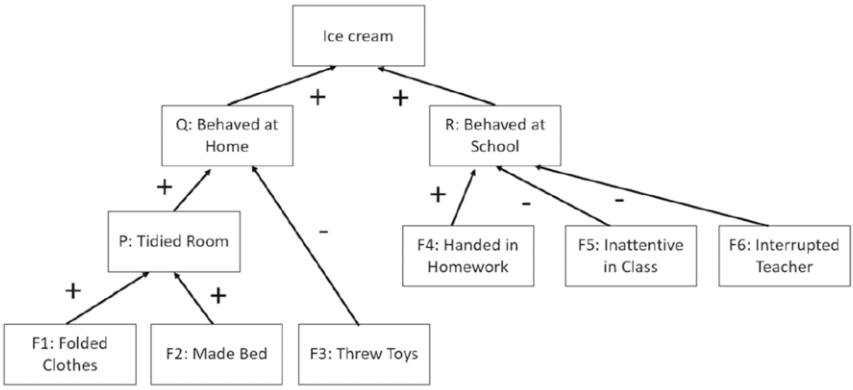



**Figure 1.** A factor hierarchy (from [3], as adapted from [4])

In Bench-Capon's first example, the case of **MaxMonday**, $F_1$ (Max had folded his clothes) and $F_5$ (Max had been inattentive in class) were present and parent Jo decided that Max could have his ice cream. Bench-Capon says that in a hierarchical model of precedential constraint there are two possible interpretations of this case.

- MMH: $Q > R$ (behaving at home is more important than behaving at school)
- MMF: $F_1 > F_5$ (folding one's clothes is more important than being inattentive in class)

Note that MMH and MMF correspond to the hierarchical and flat constraint model interpretations, respectively. According to Bench-Capon [3, p. 831], the choice between these two interpretations is not obvious. It is worthwhile quoting him in full:

> Jo may give clues as to her interpretation. If she says *ok, you tidied your room, which meant you behaved at home, so you can have ice cream even though you misbehaved at school*, she is clearly thinking in hierarchical terms. If, on the other hand she says *ok, you folded your clothes so you can have ice cream even though you were inattentive in class*, she is clearly thinking in terms of MMF.

In our opinion, which outcome is intuitively correct is relative to what is given. If the factor hierarchy is given in a way in which all base-level factors establish intermediate factors with the same strength, then there is nothing wrong with MMH. Bench-Capon's intuition can then be explained by noting that intermediate factors are not always established with the same strength. This becomes clear from his discussion of another example.

In Bench-Capon's second example, the case of **EmmaMonday**, $F_2$ (Emma had made her bed) and $F_6$ (Emma had interrupted her teacher) were present. According to

Bench-Capon, Emma would appeal to MMH and argue that Jo's decision about Max constrains Jack's decision about her to the effect that she should also have her ice cream. Bench-Capon argues that Jack could instead decide that Emma will not have her ice cream on the grounds of $F_6 > F_2$, since this preference is consistent with $F_1 > F_5$. We argue that the latter decision in fact amounts to believing that intermediate factors can in different cases be established with different strengths. Accepting both $F_1 > F_5$ and $F_6 > F_2$ in fact says that $Q$ is more strongly established by $F_1$ than by $F_2$ and that $R$ is less strongly established by $F_5$ than by $F_6$. In the next sections we will argue that the dimension-based hierarchical result model of precedential constraint proposed by [6] is suitable for modelling this approach and allows a modelling of the two examples that satisfies Bench-Capon's intuitions.

## 3. The dimension-based result model

We next summarise the dimension-based flat and hierarchical result models of precedential constraint (abbreviated as DRM and DHRM) as presented in [6]. A *dimension* is a nonempty set. Given a finite set $D$ of dimensions, a (dimension-based) *fact situation* is a partial choice function on $D$, i.e., a partial function $X : D \rightharpoonup \bigcup D$ such that $X(d) \in d$ for every $d \in D$ on which $X$ is defined. A fact situation is *complete* if it is defined on all of $D$. A *case* is a pair $(Y,s)$ where $s \in \{\pi,\delta\}$ and $Y$ is a complete dimension-based fact situation. Finally, each dimension $d$ is equipped with a partial order $\preceq$ on $d$, where $v \preceq v'$ intuitively means that $v'$ is at least as good for $\pi$ as $v$ and $v$ is at least as good for $\delta$ as $v'$.

Then 'flat' precedential constraint is defined for the DRM as follows.

**Definition 3.1.** Let $CB$ be a case base for a set $D$ of dimensions. We say that $CB$ *forces* the decision of a fact situation $X$ for $\pi$, written $CB, X \models \pi$, iff there exists a case $(Y,\pi) \in CB$ such with $Y(d) \preceq X(d)$ for all $d \in D$. Likewise, $X$ is forced for $\delta$ iff there exists a case $(Y,\pi) \in CB$ such with $X(d) \preceq Y(d)$ for all $d \in D$.

The DHRM gives the set of dimensions a hierarchical structure.

**Definition 3.2.** A *dimension hierarchy* is a pair $(D,H)$ with $D$ a finite set of dimensions and $H$ a relation on $D$ satisfying

1. the transitive closure of $H$ is irreflexive;
2. $D$ contains exactly one $H$-maximal element.

A dimension is *base-level* if it is $H$-minimal, and *abstract* otherwise. In contrast to factors, dimensions are not assumed to have an inherent polarity.[2] As such, we use the notation $H(d) = \{e \in D \mid H(e,d\}$ instead of the notations $\texttt{Pro}(d)$ and $\texttt{Con}(d)$ to refer to the direct subordinates of $d$ in the hierarchy. The case outcome is now considered part of the hierarchy, as phrased by point 2 of Definition 3.2. As such, a case is now simply defined as a complete fact situation.

Hierarchical constraint is now defined for the DHRM as follows.

**Definition 3.3.** Given a case base $CB$ and a value $v$ in some dimension $d$, a fact situation $X$ is *lower bound by* $v$ and $CB$, written $CB \models v \preceq X(d)$ iff:

[2] We make this deviation from the definitions of [6] for the sake of simplicity.

- $v \preceq X(d)$; or
- $d \in A$ and there is $Y \in CB$ satisfying $v \preceq Y(d)$ such that $CB \models Y(e) \preceq X(e)$ holds for all $e \in H(d)$.

The *upper bound by v*, written $CB \models X(d) \preceq v$, is defined analogously.

Together, the lower and upper bound of a dimension $d$ for $X$ define the range of values that $d$ can have in $X$ given the case base. As stated by [6], the idea of the recursive clause is that there is a precedent $Y$ which forces $X$ to take a value $v$ which is at least $Y(d)$, and therefore $v \preceq X(d)$ follows by transitivity.

## 4. Applying the dimension-based hierarchical result model to the examples

We now formalise our intuitions about the family example that intermediate factors can be established with varying strengths. For ease of explanation we use the natural numbers as values but our analysis applies in exactly the same way to any partially ordered set of values. For present purposes, all we need to represent is that $R$ is satisfied more strongly in *MaxMonday* than in *EmmaMonday*.

Let 'the child can (not) have ice cream' be denoted by $\pi$ ($\delta$). The factors are now dimensions but all dimensions except $P$, $Q$ and $R$ are kept two-valued as follows: we assume two values 0 and 1 where $0 \preceq 1$ for dimensions that correspond to pro-$\pi$ factors (in this case $F_1, F_2, F_4$) and $1 \preceq 0$ for dimensions that correspond to pro-$\delta$ factors (in this case $F_3, F_5, F_6$). Moreover, $P$, $Q$ and $R$ can have any natural number as value, where $v \preceq v'$ iff $v \leq v'$. Their hierarchical structure remains as in Figure 1.

We now model the intuition that making one's bed ($F_2$) makes the room more tidied ($P$) than folding one's clothes ($F_1$), and that interrupting the teacher ($F_6$) is worse behaviour at school than being inattentive in class ($F_5$). Moreover, the more the room is tidied, the better the behaviour at home ($Q$). This is modelled by assuming fact situations $M$ and $E$ which assign 0 to all dimensions except:

$$\begin{array}{ll} M\text{:} & F_1 = 1, F_5 = 1, P = 2, Q = 2, R = 3, \pi = 1 \\ E\text{:} & F_2 = 1, F_6 = 1, P = 3, Q = 3, R = ?, \pi = ? \end{array}$$

Note that we leave $E$ undefined on $R$ and $\pi$. The question is now whether the case $M$ forces a decision to grant ice cream to Emma according to the DHRM. Formally, this boils down to whether there is a lower bound $M \models 1 \preceq E(\pi)$. We show that this is not the case. Intuitively, the point is that although Emma behaved better at home than Max, she behaved worse at school than Max and the latter is a relevant difference that allows the parents to deny Emma her ice cream. Let us see how this follows formally, by unfolding Definition 3.3 with respect to the statement $M \models 1 \preceq E(\pi)$.

To have $M \models 1 \preceq E(\pi)$ (i.e., $E$ is forced to be decided for $\pi$) we must either have $1 \preceq E(\pi)$, which is not the case because $E$ is undefined on $\pi$, or else both (1) $M \models 2 \preceq E(Q)$ and (2) $M \models 3 \preceq E(R)$. The first we indeed have, since $2 \preceq 3 = E(Q)$, so let us consider (2). Successively applying Definition 3.3 and simplifying, we get:

$$M \models 3 \preceq E(R) \tag{1}$$

iff

$$3 \leq E(R), \text{ or} \tag{2}$$

$$M \models M(F_4) \preceq E(F_4), \text{ and} \tag{3}$$

$$M \models M(F_5) \preceq E(F_5), \text{ and} \tag{4}$$

$$M \models M(F_6) \preceq E(F_6) \tag{5}$$

iff

$$0 \preceq 1 \text{ (as values in } F_4), \text{ and} \tag{6}$$

$$1 \preceq 0 \text{ (as values in } F_5), \text{ and} \tag{7}$$

$$0 \preceq 1 \text{ (as values in } F_6). \tag{8}$$

In the first step, we must either have $3 \leq E(R)$, which we do not have since $E$ is undefined on $R$, or instantiate $Y$ in the second bullet of Definition 3.3 by $M$, resulting in (3)–(5). As $F_4$, $F_5$, and $F_6$ are base-level dimensions, on all of which $E$ is defined, these statements respectively simplify to (6)–(8). The last of these conjuncts, (8), does not hold: $1 \preceq 0$ in $F_6$ by definition, because interrupting the teacher is a con-$\pi$ factor. In other words, $E$ can be distinguished on $F_6$ with respect to $M$, and so $M$ does not constrain $E$ to take a value of at least 3 on $R$. We have thus verified that $M \not\models 1 \preceq E(\pi)$ according to the logic of the DHRM, so Emma can be denied her ice cream without violating precedential constraint.

We now vary the example in a way that makes Emma's case lower bounded by the decision that she is allowed to have ice cream. Consider a fact situation $E'$ which is as $E$ except that we replace $F_6$ by $F_5$, so $E'(F_5) = 1$ and $E'(F_6) = 0$. Because $M$ and $E'$ have the same values on the subordinate dimensions $F_4, F_5, F_6$ of $R$, we now do have the lower bound $M \models 3 \preceq E'(R)$. Moreover, we also have $M \models M(Q) \preceq E'(Q)$ because $M(Q) = 2 \leq E'(Q) = 3$. Putting these together, we now do have that $M \models 1 \preceq E'(\pi)$; so Emma should be allowed to have ice cream.

## 5. Conclusion

In this paper we have responded to Trevor Bench-Capon's criticism in [2,3] of hierarchical result models of precedential constraint. We have argued that his point that the models do not account for the possibility that intermediate factors are established with different strengths by different base-level factors does not apply to [6]'s dimension-based version. We have shown that in the analysis of some examples Bench-Capon seems to interpret some intermediate factors as dimensions, and that applying [6]'s hierarchical result model to these examples avoids Bench-Capon's criticisms.

## References

[1] V. Aleven. *Teaching Case-Based Argumentation Through a Model and Examples*. PhD Dissertation University of Pittsburgh, 1997.

[2] T.J.M. Bench-Capon. A note on hierarchical constraint. In G. Sileno, J. Spanakis, and G. van Dijck, editors, *Legal Knowledge and Information Systems. JURIX 2023: The Thirty-Sixth Annual Conference*, pages 3–12. IOS Press, Amsterdam etc., 2023.

[3] T.J.M. Bench-Capon. Intermediate factors and precedential constraint. *Artificial Intelligence and Law*, 33:827–846, 2025.

[4] I. Canavotto and J. Horty. Reasoning with hierarchies of open-textured predicates. In *Proceedings of the 19th International Conference on Artificial Intelligence and Law*, pages 52–61, New York, 2023. ACM Press.

[5] J. Horty. Rules and reasons in the theory of precedent. *Legal Theory*, 17:1–33, 2011.

[6] W. van Woerkom, D. Grossi, H. Prakken, and B. Verheij. Hierarchical a fortiori reasoning with dimensions. In G. Sileno, J. Spanakis, and G. van Dijck, editors, *Legal Knowledge and Information Systems. JURIX 2023: The Thirty-Sixth Annual Conference*, pages 43–52. IOS Press, Amsterdam etc., 2023.

[7] W. van Woerkom, D. Grossi, H. Prakken, and B. Verheij. Hierarchical precedential constraint. In *Proceedings of the 19th International Conference on Artificial Intelligence and Law*, pages 333–342, New York, 2023. ACM Press.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251609*

# Anomaly Detection in Legal Text – A Case Study on Cybersecurity Contracts

Pavol SOKOL [a,1], Laura BACHŇÁKOVÁ RÓZENFELDOVÁ [b], Dávid VARGA [a], Simona RUDOHRADSKÁ [b], and Regina HUČKOVÁ [b]

[a] *Faculty of Science, Pavol Jozef Šafárik University in Košice*
[b] *Faculty of Law, Pavol Jozef Šafárik University in Košice*

ORCiD ID: Pavol Sokol https://orcid.org/0000-0002-1967-8802, Laura Bachňáková Rózenfeldová https://orcid.org/0000-0002-7111-9565, Dávid Varga https://orcid.org/0000-0002-3176-8106, Simona Rudohradská https://orcid.org/0000-0002-6711-2141, Regina Hučková https://orcid.org/0000-0003-4731-0770

**Abstract.** The increasing availability of legal documentation provides opportunities for applying empirical research methods that enable a comprehensive analysis of complex legal data. Systematic Content Analysis offers a structured approach for examining legal texts. This paper examines the possibility of anomaly detection in legal text through the employment of machine learning methods. The conclusions presented are formulated on a basis of a case study focusing on contractual agreements concluded between public authorities and private service providers in the field of cybersecurity. In this regard, we propose an approach to the detection of anomalous legal text, providing a categorization of attributes that impact the determination of anomalies and describing their significance in the selected domain.

**Keywords.** systematic content analysis, machine learning, anomaly detection, legal text, cybersecurity

## 1. Introduction

Legal research focuses, primarily, on the examination of different legal documentation (legislation, case law, contractual agreements, etc.), drawing conclusions from its thorough reading and interpretation. The increasing availability of legal text provides opportunities for researchers to implement different empirical research methods enabling comprehensive analysis of complex legal data. Systematic content analysis provides a basis for such analysis, facilitating efficient examination of textual information in a standardized manner and application of different machine learning methods in this regard.

The main objective of this paper is the analysis of anomaly detection in legal texts. One of the important tasks in machine learning is the **identification of anomalies (outliers)**, and in this study we examine how this concept can be applied to contracts. Anomaly detection refers to the problem of finding patterns in data that do not conform

[1]Corresponding Author: Pavol Sokol, Institute of Computer Science, Faculty of Science, Pavol Jozef Šafárik University in Košice, Slovak Republic.

to expected behaviour, identifying significant (and often critical) actionable information in a wide variety of application domains [1]. To achieve this objective, we define the following partial research questions: (1) What constitutes an anomaly in legal texts?; (2) Which attributes contribute to anomalous contracts, and which characterize standard (non-anomalous) agreements?

To answer these questions, we focus on unsupervised outlier detection methods applied to **legal contracts in the field of cybersecurity**. The dataset consists of contracts whose subject matter is the provision of various cybersecurity-related services. In parallel, we apply **Systematic Content Analysis (SCA)**, which represents a structured empirical approach for analyzing legal texts. This dual perspective—combining outlier detection techniques with SCA allows us to identify anomalous contracts, explore attribute categories influencing their classification, and distinguish between standard contractual features and those contributing to anomalies.

The increasing employment of empirical research methods, including SCA in legal research has been discussed in numerous studies [2,3,4]. Another research group focuses on **the detection of anomalies in the judicial system and legal texts** [5,6]. The detection of anomalies in legal texts is closely related not only to the application of law but also to the legislative process. In [7], the authors focused on improving the quality of legislation and legal decisions through knowledge verification and anomaly detection.

## 2. Methodology

From a methodological point of view, this paper employs the SCA, which can be defined as a methodical, replicable technique for compressing many words of text into fewer content categories based on explicit coding rules [8]. The advantage of this approach is that it can produce both qualitative and quantitative results by enabling the researcher to compress long text into fewer content categories based on explicit rules of coding [9].

The analysis is based on a dataset of contracts obtained from the Slovak Central Register of Contracts – a public database of contracts obligatorily published pursuant to the Act No. 211/2000 Coll. on free access to information. The authors have analysed 273 contracts concluded in the time period 23.7.2019 – 1.1.2025, and used types of attributes, namely numerical (e.g., the amount of remuneration), and binary (e.g., inclusion of a contractual fine). Binary attributes represent the occurrence or the absence of a specific attribute. The analysis resulted in the selection and collection of a set of attributes classified into the six groups.

For anomaly detection we used **LSCP** [10] (Locally Selective Combination in Parallel), a dynamic ensemble that, for each sample, (1) finds a small local neighborhood (via k-nearest neighbors), (2) builds a simple pseudo-target for that neighborhood, (3) measures each base detector's competency as its correlation with that pseudo-target, and (4) averages the scores of only the most competent detectors for that sample. We ensemble six unsupervised detectors — **PCA** [11], **ECOD** [12], **COPOD** [13], **IsolationForest** [14], **HBOS** [15], and **KNN** [16]. For each sample, LSCP builds a k-NN neighborhood, forms a simple pseudo-target from neighbors' consensus, and measures each detector's competency as its correlation with that pseudo-target. These correlations are discretized into 4 bins; We set contamination to 0.05 as a standard unsupervised.

For the purpose of analyzing attributes and their impact on the anomaly of legal texts, we employed the SHapley Additive exPlanations (SHAP) [17] method. We explain LSCP

anomaly scores with *SamplingSHAP*—a faster alternative to *KernelSHAP* that preserves reliable feature-importance rankings. SHAP compares each prediction to a baseline (the model's expected output on typical cases), so we use a compact background of 30 k-means centroids to represent those cases.

## 3. Results and discussions

After running *SamplingSHAP* on the LSCP trained model, we obtained per-feature attributions that can be visualized in the beeswarm plot (Figure 1). This plot shows how all attributes contribute across the dataset, highlighting which features consistently drive anomaly detection. This explanation allows us not only to detect outliers but also to understand the underlying reasons for their anomalous character.

### *3.1. Determination of contract anomality*

A visual representation of all attributes is shown in Figure 1 in the form of a beeswarm plot. In this visualization, each dot corresponds to one (row, feature) pair. The x-axis represents the SHAP value $\phi_j(x)$, indicating how much a given feature shifts the LSCP score either upward (to the right, more anomalous) or downward (to the left, more normal). The color encodes the raw feature value for that instance (red = high value, blue = low value), while vertical jitter reflects the density of points. This plot allows us to observe which features generally increase the anomaly score (right tail) or decrease it (left tail), and whether high or low feature values are responsible for these shifts.

Contractual agreements examined in this paper and the individual attributes that characterize their content, specifically focusing on selected legal issues (definition of contractual subject matter, liability regulation, protection of confidentiality, etc.), provide a basis for the identification of 'standard' and 'atypical' contractual regulation in the selected area. The analysis of selected attributes provided in this paper identified how individual attributes and their presence (or absence) in the contractual agreements examined aided to the determination of contract anomality. In this regard, three categories of attributes could be distinguished, namely (a) attributes that add to the contract normality, (b) attributes that add to the contract anomality, and (c) attributes without a significant effect on contract classification.

The attributes, the **presence** of which in the contractual provisions is typical, and therefore which **add to contract normality** include: (1) 'Security_Solution' – confirms the inclusion of cybersecurity solution as a service in the examined contractual agreements, reflecting the general nature of contracts selected for analysis; (2) 'Fixed_Term' and 'Payment_Monthly' – the limitation of contract effectiveness is standard in contractual relations concluded between state authorities and private suppliers, as is the monthly payment of remuneration agreed between contractual parties; (3) 'Priv_Supplier_Can_Withdraw' – attribute confirming the standard protection of suppliers through the right of withdrawal from contract and (4) 'Confidentiality_Clause' – the protection of confidential information is inherent to provision of cybersecurity services. Correspondingly, the absence of the above-defined attributes adds to contract anomality.

The attributes, the **presence** of which **aids to contract anomality**, include attributes that cover (1) **services defined in the contractual subject matter**, such as the provision

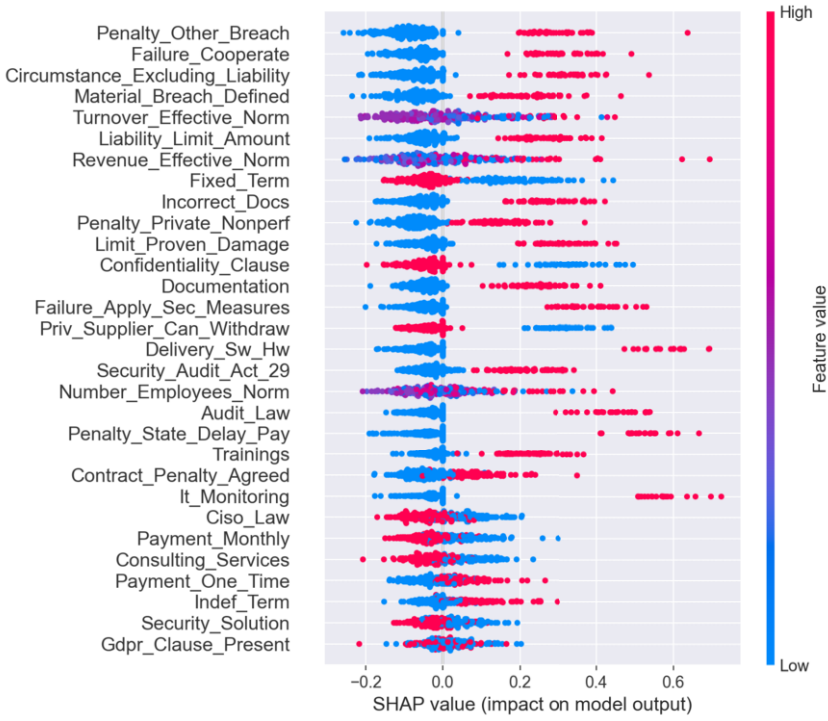



**Figure 1.** SHAP beeswarm plot for all attributes. Each dot represents one feature value, with the x-axis showing its contribution to the LSCP score (red = high value, blue = low value).

of information and cybersecurity documentation, employee training, monitoring of IT, security or external audit, (2) **the form of contract solution**, where contracts concluded for an indefinite time period or with a one-time payment are atypical for the examined contractual agreements; (3) **the limitation of liability**, where the inclusion of any provisions that restrict liability, e. g. to the specific pecuniary amount or to the amount of demonstrably caused damage, or exclude liability in defined circumstances or due to the contractual party's failure to cooperate, or act in accordance with the recommendations of the supplier, can be evaluated as contributing to the classification of contract as an anomaly, and (4) **sanctions that may be imposed in case of contract breach**, where, similarly, most attributes in this category including the stipulation of a contractual penalty in the contractual agreement as such, or inclusion of contractual penalties for the supplier's failure to perform the contract, the state's delay in remuneration payment or for the infringement of other contractual obligations, contribute to contract anomality.

The last category of attributes with no discernible impact on contract classification includes the provision of CISO (CISO_law) or cybersecurity consulting services (Consulting_services) if included in the contractual subject matter, as well as the monthly periodicity of contract renumeration payment (Payment_monthly). The dataset includes continuous attributes ('Turnover_Effective_Norm', 'Revenue_Effective_Norm', and 'Number_Employees_Norm'), the objective of which is to provide contextual information on the contractual parties, which are not included in the provisions of the contracts.

### *3.2. Discussion regarding contracts*

As part of the discussion, we focused on a specific contract that exhibits the highest anomaly (with ID 5821279), and a contract (with ID 6332291) that can be considered standard. According to statistical data obtained from the processing of selected contractual attributes, the agreement exhibiting the highest level (ID 5821279) of anomalies deviating from the norm (–0.17) concerns the provider's obligation to deliver consultancy services to the contracting authority in the areas of information security, cybersecurity, personal data protection, and IT service management. The contract shows the greatest degree of deviation among the examined attributes with respect to the imposition of a penalty on a public administration entity for delayed payment (+0.49), the delivery of incomplete documentation (+0.37), and the limitation of liability for damages (+0.3). Within the contractual framework established between the parties, the service provider is entitled, though not obliged, to claim payment of a contractual penalty amounting to 0.5% of the outstanding sum for each day of delay. Ordinarily, late payment is sanctioned by statutory or contractual default interest. The contracting parties further agreed on a limitation of liability for damages capped at the price for a "chosen period." However, the contract does not explicitly define what constitutes such a period. The total contractual price to be paid to the provider amounts to EUR 154,800, which simultaneously represents the notional ceiling of liability for damages expressed in pecuniary terms.

In contrast, the contract with the lowest level (ID 5821279) of anomalies (-0.17) refers to the consultation on cybersecurity measures and the supervision of the operator's cybersecurity and its cybersecurity manager. The least deviating attributes are the turnover criterion (–0.15), definition of material breach (–0.14), and confidentiality clauses (–0.12). The contract generally sets out grounds for withdrawal and requires both parties to protect confidential information as they would their own, promptly report any unauthorised use or disclosure, and cooperate in restoring its protection.

## 4. Conclusion

In this paper, we demonstrated a possible approach to the examination and identification of anomalies within the scope of SCA of legal text, focusing on contractual agreements concluded in cybersecurity as our case study. The proposed approach may find its application in relation to other legal text with the objective of identifying the established legal practice and possible deviations from it. The examined attributes focused on standard aspects of contractual agreements regarding their subject matter, liability, confidentiality, etc. In this regard, we calculated the impact of individual attributes on the classification of contracts as anomalous and provided their categorization into different categories, reflecting their relevance for anomaly detection. The presence of certain attributes in legal text aided in its classification as anomalous, or, correspondingly, determined the 'standard' formulation of contractual provisions in the examined case study.

## Acknowledgments

This research was carried out within the project "Automatization of legal text analysis based on machine learning" funded by the European Union Next GenerationEU through the Recovery and Resilience Plan of Slovakia under project no. 09I05-03-V02-00049.

## References


[1] Chandola V, Banerjee A, Kumar V. Anomaly Detection: A Survey. ACM Computing Surveys. 2009;41(3):1-58.

[2] Baldwin J, Davis G. Empirical Research in Law. In: Cane P, Tushnet M, editors. The Oxford Handbook of Legal Studies. Oxford: Oxford University Press; 2003. p. 880-900.

[3] Monahan J, Walker L. Twenty-Five Years of Social Science in Law. Law and Human Behavior. 2011;35(1):72-82.

[4] Leeuw FL. Empirical Legal Research: The Gap between Facts and Values and Legal Academic Training. Utrecht Law Review. 2015;11(2):19-33.

[5] Bobur M, Aibek K, Abay B, Hajiyev F. Anomaly detection between judicial text-based documents. In: 2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT). IEEE; 2020. p. 1-5.

[6] Esposito A, Di Martino B, Ammendolia R, Lupi P, Orlando M, Liang W. Time anomaly detection in the duration of civil trials in Italian justice. Connection Science. 2023;35(1):2283394. Available from: https://www.tandfonline.com/doi/pdf/10.1080/09540091.2023.2283394.

[7] Spreeuwenberg S, van Engers T, Gerrits R. The role of verification in improving the quality of legal decisionmaking. In: Legal Knowledge and Information Systems (JURIX-2001); 2001. p. 1-16.

[8] Weber RP. Basic Content Analysis. 2nd ed. Quantitative Applications in the Social Sciences. Newbury Park, CA: SAGE Publications; 1990.

[9] Burns K, Hutchinson T. The Impact of 'Empirical Facts' on Legal Scholarship and Legal Research Training. The Law Teacher. 2009;43(2):153-78.

[10] Zhao Y, Nasrullah Z, Hryniewicki MK, Li Z. LSCP: Locally Selective Combination in Parallel Outlier Ensembles. In: Proceedings of the 2019 SIAM International Conference on Data Mining (SDM). Calgary, Canada: SIAM; 2019. p. 585-93.

[11] Shyu M, Chen S, Sarinnapakorn K, Chang L. Principal Component-based Anomaly Detection Scheme. In: Lin TY, Ohsuga S, Liau C, Hu X, editors. Foundations and Novel Approaches in Data Mining. vol. 9 of Studies in Computational Intelligence. Springer; 2006. p. 311-29. Available from: https://doi.org/10.1007/11539827_18.

[12] Li Z, Zhao Y, Hu X, Botta N, Ionescu C, Chen GH. ECOD: Unsupervised Outlier Detection Using Empirical Cumulative Distribution Functions. CoRR. 2022;abs/2201.00382. Available from: https://arxiv.org/abs/2201.00382.

[13] Li Z, Zhao Y, Botta N, Ionescu C, Hu X. COPOD: Copula-Based Outlier Detection. CoRR. 2020;abs/2009.09463. Available from: https://arxiv.org/abs/2009.09463.

[14] Liu FT, Ting KM, Zhou Z. Isolation-Based Anomaly Detection. ACM Trans Knowl Discov Data. 2012;6(1):3:1-3:39. Available from: https://doi.org/10.1145/2133360.2133363.

[15] Goldstein M, Dengel A. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. KI-2012: poster and demo track. 2012;1:59-63.

[16] Angiulli F, Pizzuti C. Fast Outlier Detection in High Dimensional Spaces. In: Elomaa T, Mannila H, Toivonen H, editors. Principles of Data Mining and Knowledge Discovery, 6th European Conference, PKDD 2002, Helsinki, Finland, August 19-23, 2002, Proceedings. vol. 2431 of Lecture Notes in Computer Science. Springer; 2002. p. 15-26. Available from: https://doi.org/10.1007/3-540-45681-3_2.

[17] Lundberg SM, Lee S. A Unified Approach to Interpreting Model Predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017). Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 4768-77. Available from: https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251610*

# An Expert-Validated LLM Framework for Transforming Legal Procurement Texts into Actionable Data

Ivan SPADA [a,1], Roberto NAI [a], Davide AUDRITO [b], Vittoria Sofia Margherita TRIFILETTI [c] and Emilio SULIS [a]
[a] *University of Turin - Department of Computer Science*
[b] *University of Bologna - Department of Legal Studies*
[c] *University of Turin - Department of Law*
ORCiD ID: Ivan Spada https://orcid.org/0009-0002-0459-1189, Roberto Nai https://orcid.org/0000-0003-4031-5376, Davide Audrito https://orcid.org/0000-0002-9239-5358, Vittoria Sofia Margherita Trifiletti https://orcid.org/0009-0000-3155-758X, Emilio Sulis https://orcid.org/0000-0003-1746-3733

**Abstract.** Legal and administrative sources typically describe procedural steps that are not recorded in structured data, which makes the application of process-oriented analysis particularly challenging. In this work, we present an approach that uses Large Language Models (LLMs) to extract events and dates from unstructured legal texts. The methodology is applied to a dataset of Italian procurement notices published since 2022 on the official EU platform, Tenders Electronic Daily (TED), demonstrating how LLMs can extract valuable information, such as administrative decisions adopted prior to tender publication. These elements are incorporated into existing event logs, thereby enhancing the quality of process analysis. A sample of the extracted data has been manually reviewed by legal experts to assess the relevance and correctness of the automated detection. The results suggest that this approach can help identify procedural steps hidden in free text, thereby supporting more complete and accurate representations of legal workflows.

**Keywords.** Legal Process Analysis, Event Log Enrichment, Textual Analysis in Procurement

## 1. Introduction

The growing availability of digital data has increased attention to legal research [3]. Legal and administrative procedures, such as public procurement, are documented through legislation, notices, and rulings, yet critical process information remains in unstructured text. While information systems record standardised events, relevant procedural details often appear only in free-text sections, hindering automated analysis.

This research applies Natural Language Processing (NLP) and Large Language Models (LLMs) [25] to extract events and dates from procurement notices on the EU

[1]Corresponding Author: Ivan Spada, ivan.spada@unito.it.

platform *Tenders Electronic Daily* (TED), enriching event logs for more complete process analyses. We propose a two-stage, expert-validated framework that combines LLM-based extraction with manual legal annotation to identify procedural steps absent from unstructured data.

Our contributions are: (i) a pipeline leveraging LLMs for extracting information from procurement notices with expert supervision through dual prompting and annotation; (ii) an expert-based assessment of LLM potential and limitations in legal contexts, particularly for boolean detection tasks. Section 2 reviews related work; Section 3 presents methodology; Section 4 describes the dataset; Section 5 presents results; Section 6 concludes describing the future work.

## 2. Related Work

The automated analysis of legal procedures can be approached through business process management (BPM) [12] and, in particular, process mining (PM) [36], which reconstructs workflows from event logs to identify bottlenecks [29,30], assess compliance [31], and optimise resources [27]. However, structured event data are rarely available in legal contexts [14], where procedural information often resides in textual sources such as regulations, rulings, or administrative documents [22]. Early computational models aimed to capture the normative structure of legal processes [15], while more recent hybrid approaches combine structured data and text to improve completeness [8,26].

Information extraction has long supported the creation of legal knowledge bases through rule-based and machine learning techniques [18]. Various NLP methods have been applied to Named Entity Recognition in legal documents [2], from early statistical models [11] to domain-specific corpora for Indian [17], English [4], and German [20] legal texts. Yet, these methods remain sensitive to domain variability and linguistic ambiguity.

The advent of LLMs has brought advances in summarisation, question answering [28], and event extraction [19]. In the legal domain, LLMs have been explored for judgment prediction [37], contract analysis [32], legal summarisation [10], harmonisation [5,6], and entity–relation extraction [21]. Despite these successes, concerns about bias and explainability highlight the need for expert supervision [9].

Building on our earlier work [35,27], which introduced a process-oriented framework for event extraction from tender notices, this study extends that approach with a two-stage expert-based validation strategy, achieving higher accuracy on a larger dataset.

## 3. Methodological Framework

This section outlines the methodological framework for identifying relevant events within legal documents. Our approach integrates Large Language Models (LLMs) with expert validation to extract structured data from unstructured legal texts. The framework consists of two main stages — extraction and evaluation — supported by iterative expert feedback, as shown in Figure 1.

***Overview.*** The two-stage methodology (Figure 1) combines prompt-based event extraction with legal expert evaluation. Each stage includes an extraction phase — from prompt engineering to machine-readable linking — followed by manual expert assessment of accuracy and completeness. Two legal experts participated in the evaluation to ensure compliance with domain standards.

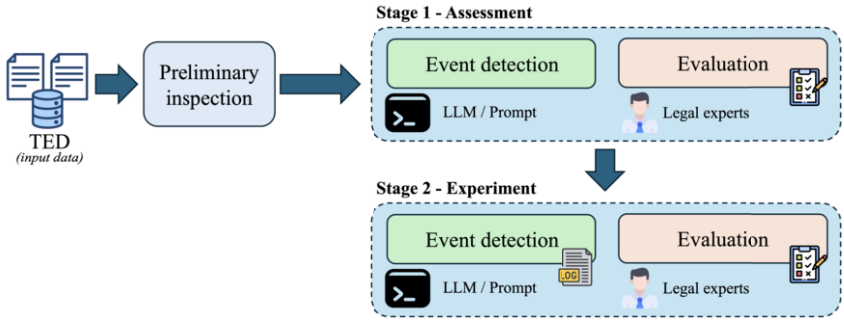



**Figure 1.** Overview of the two-stage expert-based methodology for event extraction and evaluation

***Preliminary Inspection.*** An exploratory phase was conducted with domain experts, building on previous findings [35]. This phase refined the requirements and prompts by analysing a sample of legal documents and updating the extraction guidelines. Keyword and abbreviation filtering improved document selection (e.g., DL, DM, LR, DGR), ensuring that only notices likely to contain relevant event–date pairs were included. Collaboration between legal experts and computer scientists defined both domain-specific requirements and evaluation criteria, ensuring methodological consistency.

***Stage 1.*** The assessment stage compared multiple prompting strategies across different LLMs to assess: (i) zero-shot, (ii) few-shot, and (iii) chain-of-thought (CoT) approaches. Manual evaluation of a subset of the dataset identified the best-performing model–prompt combination according to expert guidelines, enabling selection of the optimal configuration for large-scale testing.

***Stage 2.*** In the experiment stage, the selected configuration was applied to a representative sample of TED documents. Multiple experts independently annotated the results, and inter-annotator agreement was measured using Cohen's $\kappa$ [7]. This validation confirmed the reliability and domain validity of the extracted data.

## 4. From Methodology to Practice: The Case Study

This section presents the case study and its application of the proposed framework. The study analyses 27,841 Italian public procurement notices published on TED in 2022 [13]. The year 2022 was chosen because it followed the pandemic period and allowed completed procedures to be observed. While some sections of the notices follow a structured template with standardised fields, others contain free-text descriptions of procedural acts. In particular, the sections titled "complementary information" or "additional information" often include decrees and decisions marking the start of procurement procedures, which are the focus of this study.

***Preliminary inspection.*** Following our framework, we conducted an exploratory analysis of TED documents with legal experts, extending our previous work [35] that used GPT-4o [33] for event–date extraction. Legal terminology and abbreviations from a regulatory database [34] (e.g., *DL*, *DM*, *LR*, *DGR*) were integrated to refine prompts and improve document selection. The filtering pipeline progressively reduced the dataset to 2,839 relevant notices by selecting documents with "complementary" or "additional information" sections, well-formed Italian dates, and at least one legal keyword or abbreviation. This ensured both legal relevance and computational efficiency.

***Stage 1.*** Building upon [35], we extended the analysis from 2019 back to 2016 [16]. Three models were compared: GPT-4o [33], Llama-3.1-8B [24], and Claude-Sonnet-4 [1], each tested with zero-shot, few-shot, and chain-of-thought prompting. Manual evaluation on ten sample documents identified Claude-Sonnet-4 with CoT prompting as the most accurate and legally consistent configuration, particularly for handling temporal dependencies and domain-specific terminology.

***Stage 2.*** The best-performing model–prompt combination was applied to a larger sample. Two legal experts independently annotated 100 instances, achieving high inter-annotator reliability (Cohen's $\kappa$), confirming the robustness and reproducibility of the extraction process.

***Implementation details.*** The framework was implemented in Python 3.12 (64-bit) using `NLTK` for text preprocessing and APIs from OpenAI, Meta, and Anthropic. All materials—including code, guidelines, prompt strategies and datasets—are publicly available[2].

## 5. Results

***Validation of the first extraction results.*** The first extraction was evaluated by two independent annotators using a three-label scheme (Yes, No, Doubts). The overall agreement reached a Cohen's $\kappa$ of 0.576, which corresponds to a *moderate* level of inter-annotator reliability. This outcome suggests that, while the LLM-generated results are generally consistent, the presence of the Doubts category introduced some ambiguity into the annotation process, reflecting borderline or less clear-cut cases. After discussing and resolving the doubtful cases, the two legal experts proposed a final assessment of the events and dates correctly extracted by the LLM. Specifically, out of 121 cases, 89 answers were correct (74%), while the model was incorrect in 32 cases (24%). This outcome is unsatisfactory, thereby justifying a second annotation with a new prompt and a subset of notices to increase the percentage of successful cases.

***Improved annotation guidelines.*** The two-stage annotation process enhanced guideline accuracy through iterative refinement and identification of previously undetected edge cases. Legal experts revised the annotation scheme from tripartite ("Yes/No/Doubt") to binary ("Yes/No") classification. This simplification eliminated the "doubt" category, which introduced unnecessary uncertainty without discriminative value. Establishing stricter, more explicit decision criteria from the outset reduced annotator uncertainty and improved inter-annotator agreement. Additionally, the target term set was expanded

[2]GitHub repository: https://github.com/roberto-nai/JURIX-2025

to include "provision" (in Italian, "disposizione"), thereby marking events of particular relevance that had previously been excluded.

Expert guidance enabled prompt refinement by selectively filtering concept qualifiers rather than entire categories. For instance, only legislative and law decrees were excluded, as they refer to general procurement regulations rather than individual notice lifecycles, whereas other decree types remained relevant to the analysis.

***Assessment of strategies.*** The system prompt incorporates legal expert analysis of preliminary TED instances, specifying the extraction task, relevant keywords with abbreviations, and elements to exclude during event detection. This design translates legal reasoning into a computationally tractable format.

Three prompting strategies were evaluated. The zero-shot prompt (Prompt 1) provides a baseline that incorporates four expert-defined requirements, requesting JSON output with the "event" and "date" fields. The few-shot prompt (Prompt 2) extends this by adding three practical examples. The CoT prompt (Prompt 3) introduces explicit step-by-step reasoning for entity extraction in the process mining context, including a demonstration.

Manual evaluation on 10 randomly selected instances compared the three strategies across models with varying characteristics: *Claude-Sonnet-4* (Prompt 1: 8; Prompt 2: 7; Prompt 3: **9**), *GPT-4o* (Prompt 1: 7; Prompt 2: 7; Prompt 3: 8), *Llama-3.1-8B* (Prompt 1: 2; Prompt 2: 6; Prompt 3: 4). The optimal configuration was Claude-Sonnet-4 with CoT.

***Validation of second extraction results.*** Inter-annotator agreement between two legal experts on 100 cases yielded Cohen's $\kappa = 0.896$, indicating *almost perfect* agreement. The model correctly identified 108 of 115 cases (94%), with 7 errors (6%), demonstrating substantial improvement over the initial extraction and confirming the effectiveness of the expert-based framework in enhancing accuracy and reducing hallucinations. These results represent a significant advancement for legal process digitalisation and optimisation.

## 6. Conclusion and Future Work

This work presents a two-stage framework for extracting legally meaningful events and dates from administrative texts. Applied to large-scale procurement data, it enables the identification of information useful for assessing public administrations' performance and provides an empirical basis for transparency and accountability analyses. From a legal perspective, the method also helps address systemic issues such as limited transparency and corruption. During evaluation, the LLM showed a partial ability to distinguish relevant from irrelevant information—correctly ignoring non-pertinent decrees or resolutions in several cases—while still struggling with more subtle distinctions, such as those between resolutions initiating or annulling a procedure.

A key insight is that legal meaning depends on institutional context and procedural relevance rather than on isolated terms. By incorporating expert validation, the approach demonstrates that legal knowledge is essential for guiding computational analysis and producing data-driven evaluations of procurement processes.

Future work will explore the semantic modelling of procurement notices using existing ontologies and their automatic generation in queryable, explanatory [23] formats using LLMs.

## References


[1] Anthropic. Claude-Sonnet-4. https://www.anthropic.com/claude/sonnet, 2025. Accessed: 2025-09-10.

[2] Farid Ariai, Joel Mackenzie, and Gianluca Demartini. Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges. *arXiv preprint arXiv:2410.21306*, 2024.

[3] Kevin D Ashley. *Artificial intelligence and legal analytics: new tools for law practice in the digital age*. Cambridge University Press, 2017.

[4] Ting Wai Terence Au, Vasileios Lampos, and Ingemar Cox. E-NER — an annotated named entity recognition corpus of legal text. In Nikolaos Aletras, Ilias Chalkidis, Leslie Barrett, Cătălina Goanță, and Daniel Preoțiuc-Pietro, editors, *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 246–255, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[5] Davide Audrito, Luigi Di Caro, Laura Genga, Rachele Mignone, Roberto Nai, Ivan Spada, Emilio Sulis, and Vittoria Margherita Sofia Trifiletti. Standardization and harmonization of law through automated process analysis and similarity techniques. In Laura Genga, Hugo A. López, and Emilio Sulis, editors, *Proceedings of the 1st International Workshop on Processes, Laws and Compliance co-located with 6th International Conference on Process Mining (ICPM 2024), Lyngby, Denmark, October 14, 2024*, volume 3850 of *CEUR Workshop Proceedings*, pages 34–45. CEUR-WS.org, 2024.

[6] Davide Audrito, Ivan Spada, Rachele Mignone, Emilio Sulis, and Luigi Di Caro. Towards semi-automating european legislative harmonisation analysis: A harmonised glossary for llm-based legal concept detection. *Computer Law & Security Review*, 58:106171, 2025.

[7] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[8] Matilde Contestabile, Chiara Ferrara, Alberto Giovannetti, Giovanni Parrillo, and Andrea Vandin. The prolific dataset: Leveraging llms to unveil the italian lawmaking process, 2025.

[9] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93, 2024.

[10] Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. Applicability of large language models and generative models for legal case judgement summarization. *Artificial Intelligence and Law*, pages 1–44, 2024.

[11] Christopher Dozier, Ravikumar Kondadadi, Marc Light, Arun Vachher, Sriharsha Veeramachaneni, and Ramdev Wudali. Named entity recognition and resolution in legal text. In Enrico Francesconi, Simonetta Montemagni, Wim Peters, and Daniela Tiscornia, editors, *Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language*, pages 27–43. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[12] Marlon Dumas, Marcello La Rosa, Jan Mendling, and Hajo A. Reijers. *Fundamentals of Business Process Management*. Springer, 2018.

[13] European Union. Tenders electronic daily (ted) (csv subset) – public procurement notices. https://data.europa.eu/data/datasets/ted-csv?locale=en, 2024.

[14] Dirk Fahland. Extracting and pre-processing event logs. *arXiv preprint arXiv:2211.04338*, 2022.

[15] Governatori, Guido et al. Logic and the law: philosophical foundations, deontics, and defeasible reasoning. *Handbook of Deontic Logic and Normative Reasoning*, 2:655–760, 2021.

[16] Italian Government. Legislative decree no. 50/2016 and legislative decree no. 36/2023 on the public procurement code, 2023. The procurement code introduced by Legislative Decree No. 50/2016 remained in force until its replacement in March 2023 (Decree No. 36/2023).

[17] Prathamesh Kalamkar, Astha Agarwal, Aman Tiwari, Smita Gupta, Saurabh Karn, and Vivek Raghavan. Named entity recognition in Indian court judgments. In Nikolaos Aletras, Ilias Chalkidis, Leslie Barrett, Cătălina Goanță, and Daniel Preoțiuc-Pietro, editors, *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 184–193, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[18] Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. Recent advances in named entity recognition: A comprehensive survey and comparative study, 2024.

[19] Pranjal Kumar. Large language models (llms): survey, technical frameworks, and future challenges. *Artif. Intell. Rev.*, 57(9):260, 2024.

[20] Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. A dataset of German legal documents for named entity recognition. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri,

Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4478–4485, Marseille, France, May 2020. European Language Resources Association.

[21] Shiye Li and Li Yi. A few-shot entity relation extraction method in the legal domain based on large language models. In *Proceedings of the 2024 Guangdong-Hong Kong-Macao Greater Bay Area International Conference on Digital Economy and Artificial Intelligence*, DEAI '24, page 580–586, New York, NY, USA, 2024. Association for Computing Machinery.

[22] Lauren Martin, Nick Whitehouse, Stephanie Yiu, Lizzie Catterson, and Rivindu Perera. Better call gpt, comparing large language models against lawyers. *arXiv preprint arXiv:2401.16212*, 2024.

[23] Rosa Meo, Roberto Nai, and Emilio Sulis. Explainable, interpretable, trustworthy, responsible, ethical, fair, verifiable AI... what's next? In Silvia Chiusano, Tania Cerquitelli, and Robert Wrembel, editors, *Advances in Databases and Information Systems - 26th European Conference, ADBIS 2022, Turin, Italy, September 5-8, 2022, Proceedings*, volume 13389 of *Lecture Notes in Computer Science*, pages 25–34. Springer, 2022.

[24] Meta. Llama-3.1-8B. https://huggingface.co/meta-llama/Llama-3.1-8B, 2024. Accessed: 2025-09-10.

[25] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

[26] Roberto Nai, Ishrat Fatima, Gabriele Morina, Emilio Sulis, Laura Genga, Rosa Meo, and Paolo Pasteris. AI applied to the analysis of the contracts of the italian public administrations. In Fabrizio Falchi, Fosca Giannotti, Anna Monreale, Chiara Boldrini, Salvatore Rinzivillo, and Sara Colantonio, editors, *Proceedings of the Italia Intelligenza Artificiale - Thematic Workshops co-located with the 3rd CINI National Lab AIIS Conference on Artificial Intelligence (Ital IA 2023), Pisa, Italy, May 29-30, 2023*, volume 3486 of *CEUR Workshop Proceedings*, pages 255–260. CEUR-WS.org, 2023.

[27] Roberto Nai, Emilio Sulis, Davide Audrito, Vittoria Margherita Sofia Trifiletti, Rosa Meo, and Laura Genga. Leveraging process mining and event log enrichment in european public procurement analysis: a case study. *Computer Law & Security Review*, 57:106144, 2025.

[28] Roberto Nai, Emilio Sulis, Ishrat Fatima, and Rosa Meo. Large language models and recommendation systems: A proof-of-concept study on public procurements. In Amon Rapp, Luigi Di Caro, Farid Meziane, and Vijayan Sugumaran, editors, *NLDB 2024, Turin, Italy, June 25-27, 2024, Proceedings*, volume 14763 of *Lecture Notes in Computer Science*, pages 280–290. Springer, 2024.

[29] Roberto Nai, Emilio Sulis, and Laura Genga. Automated analysis with event log enrichment of the european public procurement processes. In Tiago Prince Sales, João Araújo, José Borbinha, and Giancarlo Guizzardi, editors, *Advances in Conceptual Modeling - ER 2023 Workshops, JUSMOD, Lisbon, Portugal, November 6-9, 2023, Proceedings*, volume 14319 of *Lecture Notes in Computer Science*, pages 178–188. Springer, 2023.

[30] Roberto Nai, Emilio Sulis, Rosa Meo, Francesco Gorgerino, Gabriella Margherita Racca, and Laura Genga. Process mining on a public procurement dataset: A case study. In Rosa Meo and Fabrizio Silvestri, editors, *International Workshops of ECML PKDD 2023, Turin, Italy, Sept. 18-22, 2023, Revised Selected Papers, Part I*, volume 2133 of *Communications in Computer and Information Science*, pages 477–492. Springer, 2023.

[31] Roberto Nai, Emilio Sulis, Paolo Pasteris, Mirko Giunta, and Rosa Meo. Exploitation and merge of information sources for public procurement improvement. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases - International Workshops of ECML PKDD 2022, Grenoble, France, September 19-23, 2022, Proceedings, Part I*, 2022.

[32] Savinay Narendra, Kaushal Shetty, and Adwait Ratnaparkhi. Enhancing contract negotiations with LLM-based legal document comparison. In Nikolaos Aletras, Ilias Chalkidis, Leslie Barrett, Cătălina Goanță, Daniel Preoțiuc-Pietro, and Gerasimos Spanakis, editors, *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 143–153, Miami, FL, USA, November 2024. Association for Computational Linguistics.

[33] OpenAI. GPT-4o. https://openai.com/research/gpt-4o, 2024. Accessed: 2025-09-10.

[34] PIM - Portale della Regione Lombardia. Legend of the abbreviations used in the regulatory database. https://www.pim.mi.it/legenda-dei-termini-utilizzati-nella-sezione-normativa, 2025. Accessed: 2025-10-20.

[35] Ivan Spada, Emilio Sulis, Davide Audrito, Vittoria Margherita Sofia Trifiletti, and Roberto Nai. Aug-

menting public procurement event logs with large language models: a legal process mining approach. In *Proceedings of the Joint Ontology Workshops (JOWO) - Episode XI: The Sicilian Summer under the Etna, co-located with the 15th International Conference on Formal Ontology in Information Systems (FOIS 2025)*, 2025.

[36] Wil M. P. van der Aalst. *Process Mining: Data Science in Action*. Springer, 2nd edition, 2016.

[37] Shaurya Vats, Atharva Zope, Somsubhra De, Anurag Sharma, Upal Bhattacharya, Shubham Kumar Nigam, Shouvik Kumar Guha, Koustav Rudra, and Kripabandhu Ghosh. LLMs - the good, the bad or the indispensable?: A use case on legal statute prediction and legal judgment prediction on indian court cases. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 12451–12474. Association for Computational Linguistics, 2023.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251611*

# Label Indeterminacy in AI & Law

Cor STEGING [a] Tadeusz ZBIEGIEŃ [b]
[a] *Bernoulli Institute of Mathematics, Computer Science and Artificial Intelligence, University of Groningen*
[b] *Department of Legal Theory, Jagiellonian University*

**Abstract.** Machine learning is increasingly used in the legal domain, typically operating retrospectively by treating past case outcomes as ground truth. However, legal outcomes are often shaped by human interventions that are not captured in most machine learning approaches. A final decision may result from a settlement, an appeal, or other procedural actions. This creates label indeterminacy: the outcome could have been different if the intervention had or had not taken place. We argue that legal machine learning applications need to account for label indeterminacy. Methods to impute indeterminate labels exist, but are all grounded in unverifiable assumptions. In the context of classifying cases from the European Court of Human Rights, we show that the way that labels are constructed during training can significantly affect model behaviour. We therefore position label indeterminacy as a relevant concern in AI & Law and demonstrate how it can shape model behaviour.

**Keywords.** Label indeterminacy, machine learning, legal decision-making

## 1. Introduction

With recent advances in machine learning, AI is increasingly applied in law, yet many approaches remain ill-suited for legal tasks. Limitations include the retrospective nature of machine learning, lack of faithful explanations [1,2], and, in generative models, issues such as hallucinations [3]. One crucial machine learning issue that has not yet been explored in the legal domain is *label indeterminacy* [4], where the labels of certain training instances are inherently unknown because an intervention by a decision maker has affected the final outcome. For example, in medicine, if life-sustaining treatment is withheld, it is impossible to know whether the patient would have survived if treated [5]. The resulting selectively labelled data can strongly influence model behaviour [4].

Many analogous scenarios exist in law, including bail decisions, settlements, and appeals. When bail is denied, we cannot know whether the defendant would have appeared in court if released. Similarly, if a case is settled or not appealed, the outcome that would have resulted from litigation or an appeal remains unknown. Accounting for these indeterminate labels is vital. In this paper, we examine the legal task of court case prediction. We investigate when and why decisions in non-appealed cases should be treated as indeterminate rather than ground truth, and review methods for imputing these labels. Using European Court of Human Rights case data, we show that commonly applied strategies for handling label indeterminacy can lead to substantially different model predictions.

Further details regarding this study are available at `https://arxiv.org/abs/2510.17463`.

## 2. Background

Machine learning approaches to automated legal prediction constitute an established subfield, with various models attempting to classify case outcomes across different legal domains [6,7,8,9,10]. Most prior research treats past case outcomes as fixed labels, assuming the final judgement unambiguously resolves the dispute [11,12]. This assumption is problematic under label indeterminacy, where outcomes may depend on interventions such as appeals, settlements, or other procedural decisions.

Label indeterminacy differs from *selection bias*, which arises when datasets capture only a non-representative slice of disputes [13]. Excluding settled cases would produce selection bias, while including them creates indeterminate labels, as the true court outcome is unknowable. Thus, a 'final judgement' in a dataset may not fully reflect the underlying dispute.

Label indeterminacy is also distinct from jurisprudential *legal indeterminacy*, a debate arising from realist critiques of formalism [14,15]. It argues that legal reasoning may not yield a single inevitable answer because rules are underdetermined [16,17]. Critical legal studies scholars add that outcomes are shaped by power dynamics rather than rule-following [17,18]. Dworkin [19] rebutted this, arguing that even hard cases have a right 'answer' discoverable through proper interpretation. More broadly, the notion of 'truth' in law is often socially and institutionally constituted: Patterson challenges the correspondence view, locating truth in accepted argumentative forms [20,21,22], an approach also recognized in AI & Law research [23,24,25].

While we do not settle theoretical disputes, these perspectives support viewing legal outcomes as contingent, procedural, and context-dependent. Treating all labels as fixed ground truths risks misrepresenting the phenomena machine learning models seek to capture. Our experiments show that how indeterminate labels are handled can substantially affect model behaviour, which may reinforce historical biases and ultimately undermine the reliability of AI applications in law.

## 3. European Court of Human Rights

We investigate label indeterminacy in the domain of the European Court of Human Rights (ECtHR), where we develop a machine learning model to predict case outcomes. The ECtHR hears most cases in a seven-judge Chamber, decided by majority vote [26]. In some cases, either on referral or after relinquishment, the case proceeds to the seventeen-judge Grand Chamber, whose judgment is final. Referral is discretionary; many Chamber judgments stand without review. This raises a modelling question: if a case is not referred, should the Chamber judgment be treated as definitive? Legally, Chamber rulings are binding unless overturned, yet they lack the authority of Grand Chamber judgments. From a machine learning perspective, we cannot model both simultaneously, as their decisions can conflict, for example, when a case is appealed and overturned by the Grand Chamber.

In this study, we treat Grand Chamber judgments as ground truth and regard Chamber outcomes as indeterminate. This framework allows us to investigate how different strategies for handling label indeterminacy, such as treating Chamber decisions as provisional or excluding them entirely, affect classifier behaviour. We focus on Article 6 cases

(right to a fair trial), the largest category before the Court, and frame outcome prediction as a binary classification task (violation vs. no-violation) based on the facts section, consistent with prior ECtHR outcome prediction work [27].

| ID | Name | Description | Assumption / Risk |
|---|---|---|---|
| *corr* | Correct Chamber | Includes Chamber cases without altering their labels | Assumes past Chamber decisions reflect Grand Chamber decisions |
| *obs* | Observed only | Trains only on Grand Chamber cases | Assumes missing at random; fails if appeal is outcome-related |
| $obs + ip$ | Observed + IP | Trains only on Grand Chamber cases and applies inverse propensity weights to correct for sample bias | Assumes positivity and no unobserved confounders |
| *nn* | Nearest neighbor | Imputes Chamber labels using the most similar Grand Chamber case | Assumes a valid similarity metric |
| $exp_{all}$ | All experts | Uses each expert label individually with equal weighting | Assumes all expert inputs are equally valid, including conflicting information |
| $exp_{avg}$ | Average expert | Averages multiple expert assessments into one label | Assumes the mean is meaningful; removes disagreement information |
| $exp_{max}$ | Max expert | Labels violation if at least one judge voted for it | Assumes a judge voting for violation is correct; dissenting votes disregarded |
| $exp_{min}$ | Min expert | Labels non-violation if at least one judge voted against it | Assumes a judge voting against violation is correct; other votes disregarded |
| $exp_{agr}$ | Expert agree | Keeps only cases with full expert consensus | Assumes unanimous agreement implies correctness; ambiguous cases excluded |

**Table 1.** Overview of methods used to impute the labels of Chamber cases.

## 4. Method

We use the dataset from the ECHR Open Data project [28], which contains 5,902 Chamber and 191 Grand Chamber cases pertaining to Article 6 from 1968 to 2023. We train on the facts section to predict violation vs. non-violation as determined by the Grand Chamber. To account for temporal effects in the law, we use cases before 2015 as the training set and cases from 2015 onward as the test set [29].

ECtHR cases are often long, so we use a Longformer [1] that handles up to 4,096 tokens [30]. Chamber cases average around 2,300 tokens and Grand Chamber cases around 5,600 tokens, so many exceed the limit. We lightly preprocess to remove structural noise and apply head-tail truncation for longer cases. The label distribution is heavily skewed toward violations. To mitigate this, we create seven balanced training datasets for both

[1] `https://huggingface.co/allenai/longformer-base-4096`

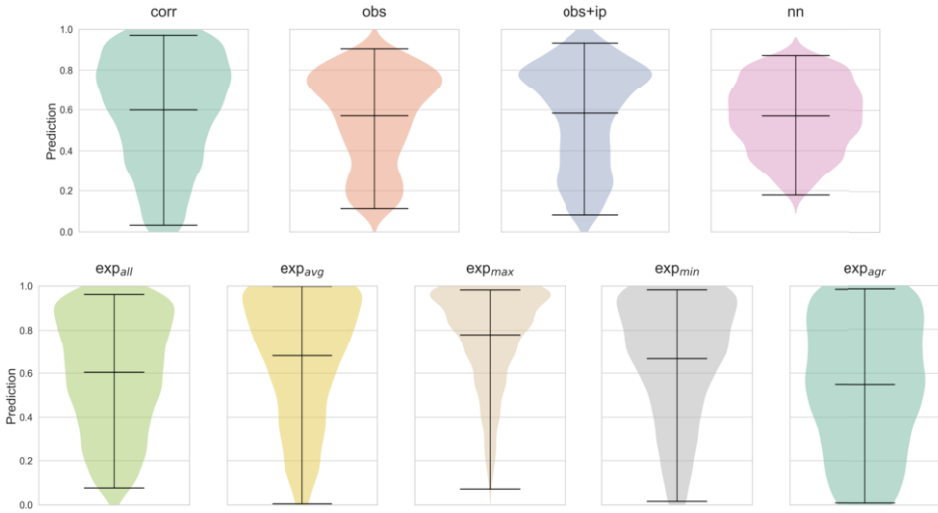
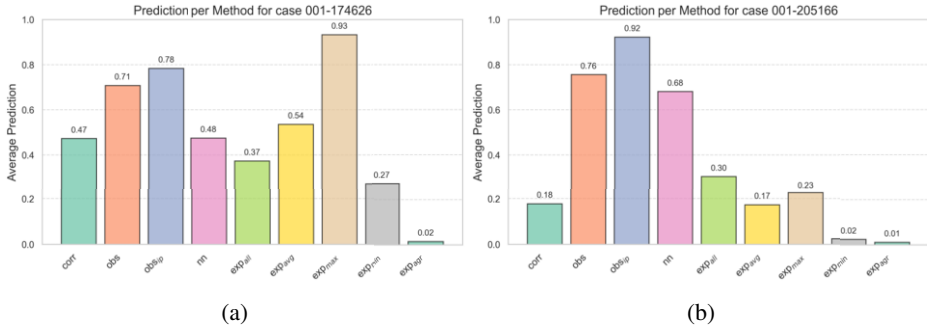



**Figure 1.** Violin plots displaying the distribution of the predictions for models trained using different label imputation methods.




(a) (b)

**Figure 2.** The mean prediction for two specific cases using different label imputation methods.

Chamber and Grand Chamber cases, each including all non-violation cases and an equal number of violation cases with minimal overlap. Test sets remain naturally imbalanced.

We model Grand Chamber decisions and treat Chamber outcomes as indeterminate, since they might have differed on appeal. To handle these indeterminate labels, we apply and compare nine imputation methods to examine their impact on model behaviour, following Schoeffer et al. [4]. Table 1 summarizes each method, with brief descriptions and the assumptions involved.

Each of the nine label imputation methods is applied to the seven balanced training sets, yielding 63 datasets for training the Longformer. We then test the model on separate Chamber and Grand Chamber test sets. For more details on model implementation, training, and full results we refer to the full paper and our GitHub repository [31]. [2]

[2] `https://github.com/CorSteging/Label-Indeterminacy-in-AI-Law`

## 5. Results

For each label imputation method, we average results across the seven balanced training sets. Figure 1 shows the distribution of predictions on both test sets, while Figure 2 illustrates the effect of different imputation methods on two specific cases.

## 6. Discussion

In this experiment, chamber case outcomes are considered indeterminate, so we cannot judge the correctness of predictions on these cases. Furthermore, all nine label imputation methods rely on unverifiable assumptions (Table 1) and none can be regarded as absolutely 'correct'. They were chosen from the literature to illustrate label indeterminacy in law, even if some assumptions may be especially implausible in legal contexts.

Figure 1 shows that predictive behaviour varies by imputation method: $exp_{max}$ generally produces high predictions near 1.0, while *nn* is more conservative. Figure 2 highlights two example cases, where predictions range up to 0.91 across methods. Examining the two cases revealed no clear reason for the divergent predictions. Both are Chamber cases, one with a single dissent and the other decided by majority, suggesting that ordinary cases such as these can be strongly affected by how label indeterminacy is handled.

Our results highlight the importance of accounting for label indeterminacy, as the choice of method can substantially affect model behaviour. Predicting ECtHR case outcomes is challenging even for humans [32], and serves mainly as an academic illustration and illustrative case study. Similar issues arise in other legal contexts, such as settlements, withdrawals, procedural defaults, bail, or appeals.

## 7. Conclusion

We introduce the concept of label indeterminacy to the legal domain and highlight its critical implications for machine learning. Certain legal outcomes are inherently unknowable because they depend on interventions whose absence or alteration could have produced different results. This creates indeterminate labels that cannot be ignored in predictive modeling. Using the ECtHR as a case study, we show that commonly applied label imputation strategies can yield different prediction behaviours, emphasizing the need to explicitly address indeterminacy when designing and evaluating AI systems in law.

## Acknowledgements

This research was partially funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl. This research was also supported by Future Law Lab under the Strategic Programme Excellence Initiative at Jagiellonian University. The authors wish to express their gratitude to Jakob Schoeffer, Piotr Bystranowski, Maciej Próchnicki, Bartosz Janik, Michał Rachalski, and Alexander Stachurski for their valuable feedback and support during the preparation of this article.

## References


[1] Bench-Capon T. The need for good old fashioned AI and Law. International Trends in Legal Informatics: A Festschrift for E Schweighofer. 2020:22-36.

[2] Steging C, Renooij S, Verheij B. Rationale discovery and explainable AI. In: Schweighofer E, editor. Legal Knowledge and Information Systems - JURIX 2021: The Thirty-Fourth Annual Conference. vol. 346 of Frontiers in Artificial Intelligence and Applications. Vilnius, Lithuania: IOS Press; 2021. p. 225-34.

[3] Dahl M, Magesh V, Suzgun M, Ho DE. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. Journal of Legal Analysis. 2024;16(1):64-93.

[4] Schoeffer J, De-Arteaga M, Elmer J. Perils of Label Indeterminacy: A Case Study on Prediction of Neurological Recovery After Cardiac Arrest. In: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency. FAccT '25. New York, NY, USA: Association for Computing Machinery; 2025. p. 1080–1094.

[5] Lakkaraju H, Kleinberg J, Leskovec J, Ludwig J, Mullainathan S. The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '17. New York, NY, USA: Association for Computing Machinery; 2017. p. 275–284.

[6] Katz DM, Bommarito MJ, Blackman J. A general approach for predicting the behavior of the Supreme Court of the United States. PLOS ONE. 2017;12(4):e0174698.

[7] Aletras N, Tsarapatsanis D, Preoţiuc-Pietro D, Lampos V. Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. PeerJ Computer Science. 2016;2:e93.

[8] Chalkidis I, Androutsopoulos I, Aletras N. Neural Legal Judgment Prediction in English. In: Korhonen A, Traum D, Màrquez L, editors. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics; 2019. p. 4317-23.

[9] Medvedeva M, Vols M, Wieling M. Using machine learning to predict decisions of the European Court of Human Rights. Artificial Intelligence and Law. 2020;28(2):237-66.

[10] Cui J, Shen X, Nie F, Wang Z, Wang J, Chen Y. A Survey on Legal Judgment Prediction: Datasets, Metrics, Models and Challenges; 2022.

[11] Galanter M, Cahill M. "Most Cases Settle": Judicial Promotion and Regulation of Settlements. Stanford Law Review. 1994;46(6):1339-91.

[12] Lever J. Why Procedure Is More Important than Substantive Law. The International and Comparative Law Quarterly. 1999;48(2):285-301.

[13] Medvedeva M, Mcbride P. Legal Judgment Prediction: If You Are Going to Do It, Do It Right. In: Preoţiuc-Pietro D, Goanta C, Chalkidis I, Barrett L, Spanakis G, Aletras N, editors. Proceedings of the Natural Legal Language Processing Workshop 2023. Singapore: Association for Computational Linguistics; 2023. p. 73-84.

[14] Hasnas J. Back to the Future: From Critical Legal Studies Forward to Legal Realism, Or How Not to Miss the Point of the Indeterminacy Argument. Duke Law Journal. 1995 Oct;45(1):84-132.

[15] Spaić B. Formalism. In: Sellers M, Kirste S, editors. Encyclopedia of the Philosophy of Law and Social Philosophy. Dordrecht: Springer; 2023. .

[16] Singer JW. The Player and the Cards: Nihilism and Legal Theory. Yale Law Journal. 1984;94(1):1.

[17] Unger RM. The Critical Legal Studies Movement. Harvard Law Review. 1983;96(3):561-675.

[18] Tushnet M. Critical Legal Studies: A Political History. The Yale Law Journal. 1991;100(5):1515-44.

[19] Dworkin R. No Right Answer? In: Hacker PMS, Raz J, editors. Law, Morality, and Society: Essays in Honour of HLA Hart. Oxford: Clarendon Press; 1977. p. 58-84.

[20] Aquinas T. Summa Theologiae. Cambridge, UK: Cambridge University Press; 2007. Original work 1265–1273, Prima Pars, Q.16.

[21] Coleman JL. Truth and objectivity in law. Legal Theory. 1995;1(1):33-68.

[22] Patterson DM. Law and Truth. New York: Oxford University Press USA; 1996.

[23] Governatori G, Bench-Capon T, Verheij B, Sartor G, Wyner A, Grabmair M, et al. Thirty years of Artificial Intelligence and Law: the first decade. Artificial Intelligence and Law. 2022;30(4):481-519.

[24] Sartor G, Araszkiewicz M, Atkinson K, Wyner A, Bench-Capon T, Governatori G, et al. Thirty years of Artificial Intelligence and Law: the second decade. Artificial Intelligence and Law. 2022;30(4):521-57.

[25] Villata S, Araszkiewicz M, Ashley K, Atkinson K, Bench-Capon T, Bex F, et al. Thirty years of Artificial Intelligence and Law: the third decade. Artificial Intelligence and Law. 2022;30(4):561-91.

[26] European Court of Human Rights. Rules of Court; 2025. Available at: https://www.echr.coe.int/documents/d/echr/rules_court_eng.
[27] Medvedeva M, Wieling M, Vols M. Rethinking the field of automatic prediction of court decisions. Artificial Intelligence and Law. 2022:1-18.
[28] Quemy A, Wrembel R. ECHR-OD: On building an integrated open repository of legal documents for machine learning applications. Information Systems. 2021:101822.
[29] Steging C, Renooij S, Verheij B. Taking the law more seriously by investigating design choices in machine learning prediction research. In: ASAIL 2023. Automated Semantic Analysis of Information in Legal Text. Proceedings of the 6th Workshop on Automated Semantic Analysis of Information in Legal Text co-located with the 19th International Conference on Artificial Intelligence and Law (ICAIL 2023). Braga, Portugal: CEUR-WS; 2023. p. 49-59.
[30] Beltagy I, Peters ME, Cohan A. Longformer: The Long-Document Transformer. arXiv:200405150. 2020.
[31] Steging C, Zbiegień T. Label Indeterminacy in AI & Law; 2025. Available from: https://arxiv.org/abs/2510.17463.
[32] Mumford J, Atkinson K, Bench-Capon T. Human Performance on the AI Legal Case Verdict Classification Task. In: Frontiers in Artificial Intelligence and Applications. vol. 379; 2023. p. 359-64.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*

*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251612*

# From Court Decisions to Guiding Principles: Advancing Complex Legal Summarization with LLMs

May Myo ZIN [a,1], Ken SATOH [a] and Georg BORGES [b]
[a] *Center for Juris-Informatics, ROIS-DS, Tokyo, Japan*
[b] *Institute of Legal Informatics, Saarland University, Saarbrücken, Germany*

**Abstract.** Guiding principles (Leitsätze) are central to German jurisprudence, capturing the essence of judicial reasoning in concise, doctrinally precise statements. Unlike general case summaries, they distill normative reasoning and key legal holdings rather than recounting factual backgrounds or procedural details. This paper examines how large language models (LLMs) can automatically generate guiding principles from German court decisions, focusing on three dimensions: model choice and adaptation, prompting strategies, and evaluation methods. Comparing GPT-4o with LerLeoLM, a fine-tuned, domain-specific model, we find that GPT-4o outperforms even without fine-tuning, while lightweight tuning on only 100 cases yields the best results. Human–LLM co-designed prompts further enhance quality, surpassing both expert-structured and self-generated prompts. Finally, we show that LLM-based evaluation aligns more closely with expert judgment than traditional metrics, establishing guiding principle generation as a distinct task in Legal NLP.

**Keywords.** large language models, guiding principles, German court decisions, prompt engineering, legal summarization, LLM-based evaluation

## 1. Introduction

Automatically generating guiding principles is a unique challenge, as it requires domain-specific abstraction, conciseness, and legal precision that go beyond the abilities of standard summarization methods, which tend to reproduce factual details or yield vague generalities. Recent large language models (LLMs) show promise in legal reasoning [1] and judgment summarization [2,3,4], but their effectiveness in generating guiding principles remains underexplored. This study investigates the use of LLMs to generate guiding principles from German court decisions. We address three core research questions:

1. **RQ1:** Can zero-shot prompting using advanced proprietary LLMs like GPT-4o rival or even surpass domain-specific models fine-tuned for this task?
2. **RQ2:** How much does expert-crafted prompting matter when models already possess strong prior knowledge of the task?
3. **RQ3:** Are conventional evaluation metrics sufficient for guiding principle generation, or is LLM-based evaluation better aligned with expert evaluation?

[1] Corresponding Author: May Myo Zin, maymyozin@nii.ac.jp

## 2. Related Work

Most research on applying LLMs to legal judgments centers on summarization tasks [6,7,8,9,16,17]. Sharma and Singh [12] used InLegalBERT [13] embeddings with K-Means clustering for extractive summarization, outperforming Legal Pegasus, T5, BART, and ChatGPT-3.5, though producing case-like summaries. Glaser et al. [14] introduced a dataset of German rulings with abstractive summaries, focusing on summarization rather than guiding principle extraction. Deroy et al. [3] compared domain-specific models and general LLMs across legal datasets. Steffes and Wiedemann [10] conceptualized the guiding principles generation task as a form of doctrinal summarization, reporting limited effectiveness from prompt-based approaches and only moderate gains from fine-tuned LLMs. Subsequent work [11] fine-tuned LLMs on judgments enriched with legal entities and proposed a class-based evaluation framework. Despite these advances, generated principles still fall short of professional quality. Overall, prior studies share key gaps: they treat guiding principle generation as summarization, rely on surface-level evaluation metrics, and face risks of hallucination [18] and doctrinal inaccuracy.

## 3. Model Selection and Fine-Tuning

We used GPT-4o as our primary advanced LLM, and *LerLeoLM* from Steffes et al. [11] as the domain-specific baseline. It is a Mistral-based model fine-tuned on 4,319 German court decisions paired with guiding principles, enriched with legal entity information following Leitner et al. [15], and represents the state of the art in German guiding principle generation. We used the publicly available outputs[2] of *LerLeoLM* on the test set, without retraining or modification. We also conducted lightweight supervised fine-tuning of GPT-4o using 100 structured training pairs. Each pair consisted of an input that combined a Prompt (with *ChatCoI-v1* shown as an example in the case box in Section 4) and a Court Decision, paired with the corresponding output, the Guiding Principle.

## 4. Prompting Strategies

As baselines, we designed three simple prompts reflecting how non-expert users might typically instruct an LLM:

- ***SimplePrompt-v1:*** *"You are a legal assistant. Write the guiding principle (headnote) for the following court decision."*
- ***SimplePrompt-v2:*** *"Read the following German court decision and formulate a Leitsatz that reflects the central legal principle."*
- ***SimplePrompt-v3:*** *"You are a legal expert preparing a Leitsatz for a case law database. From the court decision below, draft a Leitsatz (headnote) summarizing the key legal principle or doctrine derived from the ruling."*

To create the ***ExpertStructuredPrompt***, we adapted the structured prompt design proposed by Steffes et al. [10]. While it represents the most advanced legal expert-crafted prompt available, it has two main limitations: (1) its question–answer format diverges from the concise, declarative style of expert-written guiding principles, and (2) it often

[2] Available at https://github.com/bs-000/bgh-ls-llms-ler

produces overly long or additional outputs. To overcome these issues, we introduced two refinements: ***UnlabeledStructuredPrompt***, which removes the Q&A structure to align more closely with expert drafting style, and ***UnlabeledStructuredPromptCoI***, which reformulates the prompt into a chain-of-instructions (CoI) [5] format to enhance conciseness and legal fidelity.

The third prompting strategy, ***ChatCoI***, is a structured prompt composed of a chain of instructions (CoI), collaboratively created by a human and LLMs such as ChatGPT through interactive dialogue.

Interactive Dialogue for *ChatCoI* Prompt Creation

***Human:*** *What does "Leitsatz" mean in German court decisions?*
***LLM (ChatGPT-4o):*** *[Explanation of the concept of a Leitsatz.]*
***Human:*** *Yes, your understanding is correct. Now, could you suggest a chain-of-instructions (CoI) prompt for an LLM to automatically generate a Leitsatz based on a given judgment?*
***LLM (ChatGPT-4o):*** *[Proposed CoI instructions, e.g.:*
*To generate a Leitsatz, you need to:*

1. *Understand the legal context – you must know what legal issue is being decided (e.g., remonstration, liability, official duties).*
2. *Identify the court's conclusion – what did the court ultimately decide about that issue?*
3. *Extract the core legal principle – summarize that decision as a rule or principle in neutral legal language.]*

Due to space limitations, we only reproduce the shortest version of *ChatCoI* (***ChatCoI-v1***) in the paper, which is presented in the case box above. Using the same interactive dialogue in separate new chats, we also generated two additional variants: ***ChatCoI-v2*** and ***ChatCoI-v3***. A few-shot variant, ***ChatCoI-3Examples***, was also created, in which the LLM was shown three example judgment–Leitsatz pairs, instructed to learn from them, and then asked to suggest a CoI prompt.

We additionally examined self-generated CoI, ***SelfCoI***, in which the LLM autonomously generated its own task-specific instructions during inference. In this setting, the LLM was first prompted to design an internal "chain-of-instructions" for itself before generating the final *Leitsatz*. To account for variability, we executed the ***SelfCoI*** procedure three times under identical conditions.

## 5. Dataset and Evaluation Methods

We used two sources of German Federal Court of Justice (BGH) decisions. **Montagspost (2019 cases):** we collected 120 decisions issued in 2019 in the domains of criminal and administrative law from the Montagspost[3] service. Of these, 100 cases were used for GPT-4o fine-tuning and 20 cases for in-domain evaluation. **Steffes et al. evaluation dataset (2003–2022):** we additionally used the evaluation dataset of Steffes et al. [11], which contains 763 BGH decisions spanning 2003–2022, primarily in civil law. From this corpus, 20 cases were randomly selected for cross-domain evaluation.

For evaluation, we used widely adopted metrics: BLEU, ROUGE, and METEOR for lexical overlap; and BERTScore, SBERT, and GPT embeddings (*text-embedding-ada-002*) for semantic similarity. We also employed GPT-4o as a meta-evaluator. The model

[3] https://montagspost.rechtsinformatik.saarland

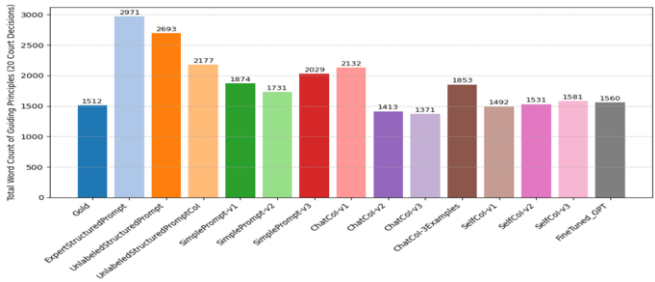



**Figure 1.** Total word counts across 20 court decisions. Each bar represents the cumulative word count of guiding principles produced using a specific prompting strategy.

**Table 1.** Comparison of different prompting and fine-tuning methods on the Montagspost (2019 cases).

| Method | BLEU-1 | BLEU-4 | METEOR | ROUGE-1 | ROUGE-L | BERT | SBERT | GPT |
|---|---|---|---|---|---|---|---|---|
| ExpertStructuredPrompt | 23.52 | 9.02 | 0.32 | 0.30 | 0.19 | 0.72 | 0.63 | 0.92 |
| UnlabeledStructuredPrompt | 25.40 | 11.07 | 0.33 | 0.32 | 0.22 | 0.73 | 0.60 | 0.92 |
| UnlabeledStructuredPromptCoI | 27.67 | 10.33 | 0.32 | 0.34 | 0.22 | 0.72 | 0.60 | 0.92 |
| SimplePrompt-v1 | 27.89 | 8.22 | 0.28 | 0.32 | 0.19 | 0.72 | 0.62 | 0.92 |
| SimplePrompt-v2 | 26.21 | 7.53 | 0.26 | 0.32 | 0.20 | 0.72 | 0.60 | 0.92 |
| SimplePrompt-v3 | 25.68 | 6.90 | 0.28 | 0.32 | 0.19 | 0.72 | 0.59 | 0.92 |
| ChatCoI-v1 | 25.80 | 8.19 | 0.27 | 0.30 | 0.18 | 0.72 | 0.60 | 0.92 |
| ChatCoI-v2 | 28.85 | 10.14 | 0.27 | 0.34 | 0.20 | 0.74 | 0.63 | 0.92 |
| ChatCoI-v3 | 26.18 | 7.32 | 0.30 | 0.32 | 0.19 | 0.72 | 0.60 | 0.92 |
| chatCoI-3Examples | 28.01 | 9.12 | 0.29 | 0.33 | 0.20 | 0.73 | 0.61 | 0.92 |
| SelfCoI-v1 | 24.22 | 5.29 | 0.24 | 0.30 | 0.17 | 0.71 | 0.59 | 0.91 |
| SelfCoI-v2 | 20.31 | 4.20 | 0.21 | 0.27 | 0.15 | 0.70 | 0.54 | 0.91 |
| SelfCoI-v3 | 25.40 | 7.52 | 0.25 | 0.30 | 0.19 | 0.71 | 0.59 | 0.91 |
| FineTuned_GPT-4o | **40.40** | **29.00** | **0.47** | **0.50** | **0.42** | **0.79** | **0.66** | **0.93** |

was prompted to act as a German legal writing expert and evaluate candidate guiding principles based on three criteria: semantic fidelity, completeness, and conciseness, each rated on a 0–5 scale. It also assigned an overall score (0–100) reflecting a holistic quality judgment rather than a sum of sub-scores. An overall score of 50 served as the threshold for acceptability. To determine whether automatic or LLM-based evaluation metrics are sufficient for assessing generated guiding principles, we additionally conducted a focused human evaluation to serve as a gold standard. Human evaluation was performed by a senior German legal scholar with expertise in civil law, legal informatics, business law, and legal theory, ensuring that judgments reflect both doctrinal accuracy and drafting conventions.

**Table 2.** Comparative evaluation of *ChatCoI-v1*-generated guiding principles against Gold references. Samples with an LLM-Evaluation Score below the threshold of 50 are highlighted in red.

| SampleID | BLEU-1 | BLEU-4 | METEOR | ROUGE-1 | ROUGE-L | BERT | SBERT | GPT | LLM-Evaluation Score (Thresh = 50) | Acceptable/ Unacceptable | Superior (Gold or Generated) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 19.82 | 6.19 | 0.36 | 0.27 | 0.17 | 0.74 | 0.51 | 0.93 | 67 | acceptable | |
| 2 | 24.27 | 11.30 | 0.34 | 0.25 | 0.13 | 0.67 | 0.27 | 0.85 | 13 | unacceptable | |
| 3 | 16.18 | 0.97 | 0.12 | 0.20 | 0.10 | 0.65 | 0.58 | 0.90 | 60 | acceptable | |
| 4 | 15.38 | 1.53 | 0.39 | 0.30 | 0.17 | 0.73 | 0.73 | 0.88 | 60 | acceptable | |
| 5 | 20.95 | 9.05 | 0.29 | 0.21 | 0.13 | 0.70 | 0.36 | 0.87 | 0 | unacceptable | |
| 6 | 21.74 | 2.28 | 0.17 | 0.24 | 0.10 | 0.71 | 0.70 | 0.96 | 67 | acceptable | |
| 7 | 29.19 | 10.43 | 0.23 | 0.33 | 0.14 | 0.68 | 0.65 | 0.91 | 60 | acceptable | Gold |
| 8 | 30.46 | 8.45 | 0.26 | 0.39 | 0.32 | 0.77 | 0.77 | 0.97 | 78 | acceptable | |
| 9 | 43.53 | 32.20 | 0.48 | 0.43 | 0.40 | 0.78 | 0.76 | 0.95 | 54 | acceptable | Generated |
| 10 | 13.45 | 4.08 | 0.15 | 0.32 | 0.18 | 0.71 | 0.65 | 0.94 | 67 | acceptable | |
| 11 | 16.59 | 4.97 | 0.16 | 0.28 | 0.16 | 0.70 | 0.68 | 0.93 | 20 | unacceptable | |
| 12 | 11.54 | 1.60 | 0.25 | 0.21 | 0.12 | 0.70 | 0.59 | 0.92 | 78 | acceptable | |
| 13 | 26.15 | 3.54 | 0.21 | 0.20 | 0.20 | 0.70 | 0.62 | 0.93 | 80 | acceptable | Generated |
| 14 | 23.42 | 3.08 | 0.20 | 0.23 | 0.16 | 0.68 | 0.43 | 0.90 | 67 | acceptable | |
| 15 | 56.60 | 24.54 | 0.59 | 0.62 | 0.39 | 0.85 | 0.66 | 0.97 | 80 | acceptable | |
| 16 | 30.00 | 4.33 | 0.19 | 0.30 | 0.12 | 0.73 | 0.65 | 0.90 | 68 | acceptable | |
| 17 | 40.18 | 16.66 | 0.31 | 0.42 | 0.21 | 0.75 | 0.69 | 0.96 | 78 | acceptable | |
| 18 | 39.83 | 15.85 | 0.39 | 0.37 | 0.25 | 0.77 | 0.77 | 0.96 | 78 | acceptable | |
| 19 | 17.95 | 1.50 | 0.18 | 0.24 | 0.11 | 0.69 | 0.48 | 0.90 | 60 | acceptable | Gold |
| 20 | 18.83 | 1.28 | 0.15 | 0.17 | 0.10 | 0.65 | 0.52 | 0.90 | 40 | unacceptable | |

**Table 3.** Comparison of different approaches on the Steffes et al. evaluation dataset (2003–2022).

| Method | BLEU-1 | BLEU-4 | METEOR | ROUGE-1 | ROUGE-L | BERT | SBERT | GPT |
|---|---|---|---|---|---|---|---|---|
| FineTuned_LerLeoLM | 16.80 | 3.71 | 0.17 | 0.22 | 0.14 | 0.68 | 0.38 | 0.89 |
| FineTuned_GPT-4o | **47.12** | **31.23** | **0.45** | **0.52** | **0.43** | **0.80** | **0.67** | **0.92** |
| SimplePrompt-v1 | 22.40 | 5.63 | 0.23 | 0.26 | 0.15 | 0.70 | 0.49 | 0.90 |
| SimplePrompt-v2 | 22.74 | 5.38 | 0.22 | 0.26 | 0.14 | 0.70 | 0.51 | 0.91 |
| SimplePrompt-v3 | 20.73 | 4.68 | 0.24 | 0.25 | 0.15 | 0.70 | 0.51 | 0.91 |
| ChatCol-v1 | 25.12 | 7.43 | 0.28 | 0.30 | 0.20 | 0.72 | 0.54 | 0.91 |
| ChatCol-v2 | 24.62 | 6.20 | 0.29 | 0.29 | 0.17 | 0.71 | 0.55 | 0.92 |
| ChatCol-v3 | 26.19 | 6.46 | 0.25 | 0.28 | 0.17 | 0.71 | 0.55 | 0.91 |

## 6. Experimental Results and Discussion

As shown in Figure1, ***fine-tuned GPT-4o*** produced guiding principles whose lengths most closely matched the "Gold" references. In contrast, both ***ExpertStructuredPrompt*** and ***UnlabeledStructuredPrompt***, which rely on expert-crafted structure, yielded substantially longer outputs, indicating that heavy structuring can encourage verbosity. Since conciseness is a defining feature of German Leitsätze, similarity in length can serve as a partial indicator of stylistic adequacy, supporting the view that overly rigid expert prompting may undermine brevity and precision, thus addressing **RQ2**.

Table 1 shows that ***fine-tuned GPT-4o*** outperforms all other methods across lexical and semantic metrics. Among prompt-only approaches, ***ChatCoI-v2*** performs best overall. To address **RQ3**, Table 2 compares automatic metrics with expert assessments. The results indicate that traditional evaluation metrics are insufficient for assessing the adequacy of generated guiding principles. Some outputs scored high on these metrics but were judged unacceptable by human experts, while others with modest scores were considered acceptable. LLM-based evaluation, using ***GPT-4o*** as a meta-evaluator, showed stronger alignment with expert judgment, making it a more reliable method for legal quality assessment. Table 3 presents results based on the Steffes et al. dataset to address **RQ1**, examining whether advanced proprietary models can outperform domain-specific fine-tuned baselines. ***GPT-4o***, even without any task-specific training, outperformed the ***fine-tuned LerLeoLM*** across all evaluation metrics. Notably, when ***GPT-4o*** was fine-tuned on just 100 examples (drawn from 2019 criminal and administrative cases), it achieved substantial performance gains despite the domain and temporal mismatch.

Overall, ***fine-tuned GPT-4o*** delivers state-of-the-art performance. Among non-fine-tuned approaches, ***ChatCoI*** prompts stand out for their robustness. Even without expert knowledge, users can co-create effective prompts through interactive prompting with an LLM. This highlights the practical potential of collaborative prompting in real-world legal NLP tasks. All prompts used in this study are available in the online repository.

## 7. Conclusion

In this paper, we present the first systematic investigation of guiding principle generation using LLMs. Our results show that GPT-4o surpasses a domain-specific baseline even with minimal and simple prompting, and that human–LLM collaborative prompt creation yields more doctrinally accurate outputs than expert-designed strategies. At the same time, we highlight the limitations of surface-level metrics and propose LLM-based evaluation as a scalable complement to expert review. While our study is limited by dataset size and jurisdiction, it establishes guiding principle generation as a distinct task that requires abstraction beyond summarization and opens new directions for research on doctrinally faithful legal NLP.

## Acknowledgements

This research was supported by the "Strategic Research Projects" grant from ROIS (Research Organization of Information and Systems), the "R&D Hub Aimed at Ensuring Transparency and Reliability of Generative AI Models" project of the Ministry of Education, Culture, Sports, Science and Technology, and JSPS KAKENHI Grant Numbers JP22H00543, JP25H00522, JP25H01112 and JP25H01152.

## References

[1] Yuan W, Cao J, Jiang Z, Kang Y, Lin J, Song K, Yan P, Sun C, Liu X. Can large language models grasp legal theories? enhance legal reasoning with insights from multi-agent collaboration. arXiv preprint arXiv:2410.02507. 2024 Oct 3.

[2] Xu H, Savelka J, Ashley KD. Toward summarizing case decisions via extracting argument issues, reasons, and conclusions. InProceedings of the eighteenth international conference on artificial intelligence and law 2021 Jun 21 (pp. 250-254).

[3] Deroy A, Ghosh K, Ghosh S. Applicability of large language models and generative models for legal case judgement summarization. Artificial Intelligence and Law. 2024 Jul 15:1-44.

[4] Kumar H, Jayanth P. Large Language Models for Indian Legal Text Summarisation. In2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT) 2024 Jul 12 (pp. 1-5). IEEE.

[5] Zin MM, Satoh K, Borges G. Leveraging LLM for Identification and Extraction of Normative Statements. InLegal Knowledge and Information Systems 2024 (pp. 215-225). IOS Press.

[6] Schubiger R.German summarization with large language models (Master's thesis, ETH Zurich).

[7] Deroy A, Ghosh K, Ghosh S. How ready are pre-trained abstractive models and LLMs for legal case judgement summarization?. arXiv preprint arXiv:2306.01248. 2023 Jun 2.

[8] Liu S, Cao J, Li Y, Yang R, Wen Z. Low-resource court judgment summarization for common law systems. Information Processing & Management. 2024 Sep 1;61(5):103796.

[9] Banerjee P, Mazumdar A, Ansar W, Goswami S, Chakrabarti A. A Data Science Approach to Calcutta High Court Judgments: An Efficient LLM and RAG-powered Framework for Summarization and Similar Cases Retrieval. arXiv preprint arXiv:2507.01058. 2025 Jun 28.

[10] Steffes B, Wiedemann NT. Generating Guiding Principles: Evaluating Large Language Models for Complex German Legal Summaries. InJSAI International Symposium on Artificial Intelligence 2025 May 23 (pp. 49-65). Singapore: Springer Nature Singapore.

[11] Steffes B, Wiedemann NT, Gratz A, Hochreither P, Meyer JE, Schilke KL. Summarisation of German Judgments in conjunction with a Class-based Evaluation. arXiv preprint arXiv:2505.05947. 2025 May 9.

[12] Sharma S, Singh PP. Advancements in legal text summarization: integrating InLegalBERT for effective extractive summarization. International Journal of System Assurance Engineering and Management. 2025 Apr;16(4):1382-97.

[13] Paul S, Mandal A, Goyal P, Ghosh S. Pre-trained language models for the legal domain: a case study on Indian law. InProceedings of the nineteenth international conference on artificial intelligence and law 2023 Jun 19 (pp. 187-196).

[14] Glaser I, Moser S, Matthes F. Summarization of German court rulings. InProceedings of the Natural Legal Language Processing Workshop 2021 2021 Nov (pp. 180-189).

[15] Leitner E, Rehm G, Moreno-Schneider J. A dataset of german legal documents for named entity recognition. arxiv 2020. arXiv preprint arXiv:2003.13016.

[16] Akter M, Çano E, Weber E, Dobler D, Habernal I. A Comprehensive Survey on Legal Summarization: Challenges and Future Directions. arXiv preprint arXiv:2501.17830. 2025 Jan 29.

[17] Ajay Mukund S, Easwarakumar KS. Optimizing Legal Text Summarization Through Dynamic Retrieval-Augmented Generation and Domain-Specific Adaptation. Symmetry. 2025 Apr 23;17(5):633.

[18] Magesh V, Surani F, Dahl M, Suzgun M, Manning CD, Ho DE. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. Journal of Empirical Legal Studies. 2025 Jun;22(2):216-42.

# Posters

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251614*

# LR$^2$: A Legal RAG for Legal Reasoning over Cases

Irene BENEDETTO [a,b,1] Luca CAGLIERO [b] and Francesco TARASCONI [c]
[a] *JAKALA, Via Borsellino, 17, Torino, 10139, Italy*
[b] *Politecnico di Torino, Corso Castelfidardo, 39, Torino, 10129, Italy*
[c] *Aruba SPA, Via San Clemente 53, Ponte San Pietro, Italy*
ORCiD ID: Irene Benedetto https://orcid.org/0000-0001-7086-7898, Luca Cagliero https://orcid.org/0000-0002-7185-5247, Francesco Tarasconi https://orcid.org/0000-0003-4562-2463

**Abstract.** This paper presents a Retrieval Augmented Generation (RAG) pipeline adapted to simulate the process of resolving legal cases. It introduces two novel approaches to enhance the RAG pipeline for legal reasoning: 1) *problem generation*, which automatically creates fictitious legal cases to improve retrieval performance, and 2) *query rewriting*, which reformulates user queries to better align with relevant passages. These methods are evaluated on a legal reasoning dataset derived from real U.S. civil procedure cases. Experiments conducted across different retrieval methods (keyword search, tensor search, and hybrid search with re-ranking) and language models (GPT-4, SAULM, Llama-3 8B) highlight the benefits of the more advanced RAG pipeline. Notably, the query rewriting approach consistently improves performance by enhancing the alignment between queries and passages.

**Keywords.** Retrieval Augmented Generation, Legal Reasoning

## 1. Introduction

Artificial Intelligence has significant potential to assist legal professionals by reducing the time and effort required for complex legal tasks. Recent research has explored LLMs' capabilities in legal reasoning through prompting-based approaches for tasks such as legal judgment prediction [1], statutory reasoning [2], and legal exams [3], though limited attention has been given to open-source models [4] or their integration with closed-source systems like GPT-4 [5]. RAG systems have evolved from basic retrieve-read pipelines [6] to advanced modular architectures with adaptive retrieval flows [7–9], though these sophisticated approaches often require substantial computational resources and data [10, 11]. In the legal domain specifically, RAG-based systems have been developed for statutory law inquiries in different languages [12, 13], highlighting the need for more comprehensive approaches to legal question-answering systems.

This paper addresses the task of legal reasoning over court cases, which involves a two-step process: (1) retrieving the most pertinent passages from case text, such as statutory provisions, precedents, or doctrinal explications, and (2) analyzing the retrieved materials to determine the most appropriate resolution. It introduces LR$^2$ (Legal RAG

[1] Corresponding Author: Irene Benedetto, irene.benedetto@jakala.com

for Legal Reasoning), a RAG pipeline specifically designed to address query-passage misalignment issues that negatively impact retrieval performance in legal contexts.

## 2. Methodology

The proposed LR² pipeline builds upon the standard Retrieve-Read RAG architecture [6] with two novel enhancements to address query-passage misalignment:
*Problem Generation*: Instead of indexing original passages directly, the system automatically generates fictitious legal cases from each original passage using an LLM. For each passage $p_j$, a corresponding fictitious legal problem $p'_j$ is generated. The retriever then computes similarity scores $s(q_i, p'_j)$ between the user query and generated problems, ranking passages and returning the top-k original passages $p_j$ associated with the top-scores generated problems. This approach aims to better align the knowledge base content with typical user queries.
*Query Rewriting*: An LLM reformulates the user input $q_i$ to extract the most important terms, producing a refined query $q'_i$. The retrieval module then computes scores $s(q'_i, p_j)$ for each original passage, while the generation component uses the original user input $q_i$ with the retrieved passages to produce the final answer. This method enhances retrieval effectiveness by improving query-passage alignment.

The dataset used for evaluation is derived from U.S. civil procedure domain [14, 15], converted into a multiple-choice question answering format using Multiple-Choice Prompting (MCP) [16]. Each instance consists of a legal problem question, context introduction, possible solution arguments, and annotated labels indicating correctness.

## 3. Results

GPT-4.1 demonstrated the best overall performance across all methods, with query rewriting achieving the highest scores (Accuracy: 0.632, Precision: 0.620, Recall: 0.611, F1: 0.613). For GPT-4, query rewriting yielded substantial improvements (Accuracy: 0.590, F1: 0.550) compared to the baseline Retrieve-Read approach (Accuracy: 0.515, F1: 0.414). Llama-3 8B showed modest improvements with query rewriting (Accuracy: 0.410, F1: 0.402) over baseline (Accuracy: 0.401, F1: 0.382). SAULM, despite being pre-trained on legal domain, demonstrated lower generalization capability in zero-shot reasoning tasks. The hybrid BM25 with tensor re-ranking consistently outperformed other methods. Query rewriting with this hybrid approach achieved the best retrieval metrics (MRR: 0.368, R@1: 0.234, R@5: 0.563, R@10: 0.688), outperforming the baseline Retrieve-Read method (MRR: 0.337, R@1: 0.219, R@10: 0.633). Problem generation yielded modest improvements (MRR: 0.359, R@1: 0.211) but did not match query rewriting's effectiveness. Methods relying solely on dense vector representations consistently underperformed compared to keyword-based and hybrid approaches, suggesting that explicit keyword matching remains crucial for legal information retrieval.

## 4. Acknowledgements

This study has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101093026 (project EFRA) and was partly carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PNRR – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013).

## References


[1] Jiang C, Yang X. Legal Syllogism Prompting: Teaching Large Language Models for Legal Judgment Prediction; 2023.

[2] Blair-Stanek A, Holzenberger N, Durme BV. Can GPT-3 Perform Statutory Reasoning?; 2023.

[3] Yu F, Quartey L, Schilder F. Exploring the Effectiveness of Prompt Engineering for Legal Reasoning Tasks. In: Rogers A, Boyd-Graber J, Okazaki N, editors. Findings of the Association for Computational Linguistics: ACL 2023. Toronto, Canada: Association for Computational Linguistics; 2023. p. 13582-96. Available from: https://aclanthology.org/2023.findings-acl.858.

[4] Guha N, Nyarko J, Ho DE, Ré C, Chilton A, Narayana A, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. arXiv preprint arXiv:230811462. 2023.

[5] OpenAI, :, Achiam J, Adler S, Agarwal S, Ahmad L, et al.. GPT-4 Technical Report; 2023.

[6] Ma X, Gong Y, He P, Zhao H, Duan N. Query Rewriting in Retrieval-Augmented Large Language Models. In: Bouamor H, Pino J, Bali K, editors. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics; 2023. p. 5303-15. Available from: https://aclanthology.org/2023.emnlp-main.322.

[7] Shao Z, Gong Y, Shen Y, Huang M, Duan N, Chen W. Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy; 2023. Available from: https://arxiv.org/abs/2305.15294.

[8] Jiang Z, Xu FF, Gao L, Sun Z, Liu Q, Dwivedi-Yu J, et al.. Active Retrieval Augmented Generation; 2023. Available from: https://arxiv.org/abs/2305.06983.

[9] Asai A, Wu Z, Wang Y, Sil A, Hajishirzi H. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection; 2023. Available from: https://arxiv.org/abs/2310.11511.

[10] Yang H, Li Z, Zhang Y, Wang J, Cheng N, Li M, et al. PRCA: Fitting Black-Box Large Language Models for Retrieval Question Answering via Pluggable Reward-Driven Contextual Adapter. In: Bouamor H, Pino J, Bali K, editors. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics; 2023. p. 5364-75. Available from: https://aclanthology.org/2023.emnlp-main.326.

[11] Yu Z, Xiong C, Yu S, Liu Z. Augmentation-Adapted Retriever Improves Generalization of Language Models as Generic Plug-In; 2023. Available from: https://arxiv.org/abs/2305.17331.

[12] Louis A, van Dijck G, Spanakis G. Interpretable Long-Form Legal Question Answering with Retrieval-Augmented Large Language Models. In: AAAI Conference on Artificial Intelligence; 2023. Available from: https://api.semanticscholar.org/CorpusID:263310713.

[13] Wiratunga N, Abeyratne R, Jayawardena L, Martin K, Massie S, Nkisi-Orji I, et al.. CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering; 2024.

[14] Held L, Habernal I. SemEval-2024 Task 5: Argument Reasoning in Civil Procedure. In: Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024). Mexico City, Mexico: Association for Computational Linguistics; 2024. .

[15] Bongard L, Held L, Habernal I. The Legal Argument Reasoning Task in Civil Procedure. In: Aletras N, Chalkidis I, Barrett L, Goanță C, Preoțiuc-Pietro D, editors. Proceedings of the Natural Legal Language Processing Workshop 2022. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics; 2022. p. 194-207. Available from: https://aclanthology.org/2022.nllp-1.17.

[16] Robinson J, Rytting CM, Wingate D. Leveraging Large Language Models for Multiple Choice Question Answering; 2023.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251615*

# Discovering the Potential of LLMs in Annotating Legal Texts for Argument Mining (Extended Abstract)

Christina BERGHEGGER [a], César PHILIPPE [a], Karla SALAS-JIMENEZ [a,1], Jean-Guy MAILLY [a], Leila MOUDJARI [a], Laurent PERRUSSEL [a]

[a] *IRIT, Université Toulouse Capitole*

**Abstract.** Argument Mining (AM) [1], the task of extracting arguments and their relations (*e.g.* attacks, supports) from text, has lead to potential applications of formal models of argumentation on real-world scenarios. However, AM methods largely depend on Machine Learning (ML) techniques, which require high-quality annotated data. This poses a significant limitation in domains like legal reasoning, where annotated corpora are extremely scarce. Current legal AM research relies almost entirely on a single dataset from the European Court of Human Rights (ECHR) [2], limiting the development of robust, argumentation-driven automated legal reasoning systems. Addressing this data scarcity is essential for advancing computational legal argumentation. To address this challenge, this paper briefly describes some preliminary results that demonstrate the promising potential of LLMs to assist legal experts in annotating additional legal corpora. This advancement paves the way for more robust, argumentation-driven automated reasoning systems in the legal domain. More details on our approach and experimental results are provided in [3].

**Keywords.** Legal argumentation, Argument mining, LLMs

## 1. Argumentation Pipeline

Using 9 short and 6 long texts from the ECHR dataset as a baseline, we have queried an LLM (*meta-llama/Meta-Llama-3-8B-Instruct* [4]) to extract arguments from these texts, aiming to identify how many original (annotated) arguments are generated by the model. Since LLMs may rephrase the content, we measured the similarity between original and generated arguments using the *sentence-transformers/all-MiniLM-L6-v2* similarity model. A generated argument is considered a match if its similarity to an original one exceeds a threshold $\tau$. We tested different pipeline variants (outlined in Figure 1), with or without providing examples of arguments.

[1] Corresponding Author: Karla-Denia.Salas-Jimenez@irit.fr.

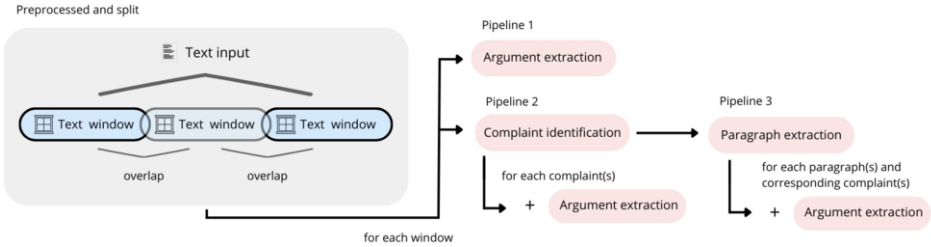



**Figure 1.** Schematic description of the three variants of our pipeline, where the text input is split in several text windows, and then either (1) the LLM is used to extract arguments directly, or (2) the complaints in the windows are identified before the argument extraction, or (3) for each complaint, there is a paragraph extraction before argument extraction.

## 2. Experiments

Our initial experiments show that the first pipeline (direct extraction of arguments from legal texts using examples) is the most effective. The number of arguments generated by the LLM and their mean and maximal number of premises are the most similar to those in the annotated dataset. So, we focus on this pipeline to evaluate similarity and matching ratios between original and LLM-generated arguments. After a first set of experiments, we introduced an improved prompt, which yields better results on two text samples (short and long texts). Although matched arguments generally show high similarity (over 80% on average), many generated arguments do not appear in the original dataset. Our improved prompt allows to reduce this phenomenon. Results are shown in Table 1, for short and long texts, before and after prompt modification.

| Based on | 0-shot | before | after | long-texts before | long-texts after |
|---|---|---|---|---|---|
| Original | 70.69 | 64.85 | 76.40 | 70.79 | 84.81 |
| LLM | 55.16 | 62.44 | 61.92 | 41.01 | 46.47 |

**Table 1.** Mean matching ratio, $\tau = 0.75$, for all prompt modifications. The first line refers to the values of the original arguments and the second to the values of the LLM arguments.

## 3. Conclusion

The evaluation of our system shows promising capabilities for LLMs to help identifying arguments in legal text, which would greatly reduce the burden of annotation. Besides the results presented in this extended abstract, we have also evaluated another LLM (*Equall/Saul-7B-Instruct-v1* [5]) and a state-of-the-art AM approach (ArgueMapper/ArgueBuf [6,7]), as well as the similarity between claims and premises of the extracted arguments. As part of our future work, we plan to evaluate the efficiency of our proposed pipeline as a support tool for legal experts in the annotation of legal texts. Additionally, we aim to assess its versatility by extending its application beyond the ECHR dataset, with the goal of generalizing the approach to a broader range of legal domains.

## Acknowledgements

This work has been funded by the French National Research Agency (grant AIDAL, ANR-22-CPJ1-0061-01).

## References

[1] Cabrio E, Villata S. Five Years of Argument Mining: a Data-driven Analysis. In: Lang J, editor. Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden. ijcai.org; 2018. p. 5427-33. Available from: https://doi.org/10.24963/ijcai.2018/766.

[2] Poudyal P, Savelka J, Ieven A, Moens MF, Goncalves T, Quaresma P. ECHR: Legal Corpus for Argument Mining. In: Cabrio E, Villata S, editors. Proceedings of the 7th Workshop on Argument Mining. Online: Association for Computational Linguistics; 2020. p. 67-75. Available from: https://aclanthology.org/2020.argmining-1.8/.

[3] Berghegger C, Philippe C, Salas-Jimenez K, Mailly JG, Moudjari L, Perrussel L. Discovering the Potential of LLMs in Annotating Legal Texts for Argument Mining. In: Second International Workshop on Argumentation and Applications (Arg&App 2025); 2025. To appear.

[4] AI@Meta. Llama 3 Model Card; 2024. Available from: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[5] Colombo P, Pires TP, Boudiaf M, Culver D, Melo R, Corro C, et al.. SaulLM-7B: A pioneering Large Language Model for Law; 2024. Available from: https://arxiv.org/abs/2403.03883.

[6] Lenz M, Sahitaj P, Kallenberg S, Coors C, Dumani L, Schenkel R, et al. Towards an Argument Mining Pipeline Transforming Texts to Argument Graphs. In: Prakken H, Bistarelli S, Santini F, Taticchi C, editors. Computational Models of Argument - Proceedings of COMMA 2020, Perugia, Italy, September 4-11, 2020. vol. 326 of Frontiers in Artificial Intelligence and Applications. IOS Press; 2020. p. 263-70. Available from: https://doi.org/10.3233/FAIA200510.

[7] Lenz M, Bergmann R. User-Centric Argument Mining with ArgueMapper and Arguebuf. In: Toni F, Polberg S, Booth R, Caminada M, Kido H, editors. Computational Models of Argument - Proceedings of COMMA 2022, Cardiff, Wales, UK, 14-16 September 2022. vol. 353 of Frontiers in Artificial Intelligence and Applications. IOS Press; 2022. p. 367-8. Available from: https://doi.org/10.3233/FAIA220176.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*

*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251616*

# Pro-Judice: Aligning LLMs with Procedural Fairness in Judicial Contexts

Qingjing CHEN [a,1] Rongxin CHENG [b] Yan LIU [b] Siyuan ZHENG [c]
Antonino ROTOLO [a,2] Yun LIU [b] Weixing SHEN [b]
[a] *Alma AI, Faculty of Law, University of Bologna, Italy*
[b] *School of Law, Institute for AI and Law, Tsinghua University, China*
[c] *Shanghai Jiao Tong University, China*

**Abstract.** This study introduces *Pro-Judice*, a benchmark to assess how large language models align with procedural fairness in judicial contexts. Using criminal case data from China and US legal traditions, it reveals that models' procedural fairness perceptions vary by legal context of test dataset, model architecture and version differences, and dataset-model interaction.

**Keywords.** Large Language Models (LLMs), Procedural Fairness, Judicial Alignment, Legal Tradition

## 1. Introduction

In legal AI, judgment prediction is a key focus, but concerns remain due to the black-box nature of large language models. [1][2][3][4][5]. While prior work emphasizes human-computer decision making and distributive justice, procedural fairness is underexplored.[6][7][8][9][10][11] [12] [13][14]Procedures are vital for legitimacy and rights protection, especially in criminal law. Legal traditions differ: civil law stresses fact-finding, common law values adversarial fairness.[15] AI models may inherit such biases from training data.[16][17][18]This study introduces the *Pro-Judice* benchmark, grounded in Rawls' theory[19] and international human rights standards, using criminal law data from China and the U.S. It evaluates LLMs' alignment with nine procedural fairness principles, evaluating five models—DeepSeek V3, DeepSeek R1, and Qwen-2.5 from China, and GPT-4o and Llama-3.3 from the U.S. Results show that fairness perceptions vary by legal tradition of model and dataset, and dataset-model interaction.

This study ran three experiments using procedural fairness means as metrics, and applied Statistical Significance Testing (Table 1) to compare models, datasets, and their interaction effects.E1 evaluates the mean of Procedural Fairness Awareness MPA $= N_{\text{no}}/N_{\text{total}}$. LLMs identify procedural unfairness with 80% accuracy. Evaluation on U.S. datasets shows stronger procedural awareness. DeepSeek R1 performed significantly worse than V3. E2 introduces MPV $= (N_{\text{pro}} - N_{\text{sub}})/N_{\text{total}}$, to assess the model's prefer-

[1]The first, second, and third authors contributed equally to this research.
[2]Corresponding Author: Antonino Rotolo, antonino.rotolo@unibo.it; Yun Liu, liuyun89@tsinghua.edu.cn; Weixing Shen, wxshen@mail.tsinghua.edu.cn. Dataset and code: https://github.com/CrexCheng/Pro-Judice

| Dimension | Group | | E1 | | E2 | | E3 | |
|---|---|---|---|---|---|---|---|---|
| | Group2 | Group1 | $\Delta M_{PA}(G2-G1)$ | P | $\Delta M_{PV}(G2-G1)$ | P | $\Delta M_{PE}(G2-G1)$ | P |
| **Dataset(LT)** | $D_{US}$ | $D_{CN}$ | 0.100 | <0.001*** | 0.841 | <0.001*** | 22.341 | <0.001*** |
| **Model(Ver)** | $M_{R1}$ | $M_{V3}$ | -0.231 | <0.001*** | -0.004 | 0.799 | 10.094 | <0.001*** |
| **Model(LT)** | $M_{US}$ | $M_{CN}$ | -0.022 | 0.057 | -0.037 | <0.001*** | 8.500 | <0.001*** |
| **Dataset x model(Ver)** | $D_{US}(R1)$ | $D_{US}(V3)$ | -0.173 | <0.001*** | 0.371 | <0.001*** | 17.358 | <0.001*** |
| | $D_{CN}(R1)$ | $D_{CN}(V3)$ | -0.290 | | -0.378 | | 2.753 | |
| **Dataset x model(LT)** | $D_{US}(M_{US})$ | $D_{US}(M_{CN})$ | -0.054 | <0.001*** | -0.160 | <0.001*** | 14.696 | <0.001*** |
| | $D_{CN}(M_{US})$ | $D_{CN}(M_{CN})$ | 0.011 | | 0.086 | | 2.302 | |

**Table 1.** Statistical Significance Testing across models, datasetst-tests and their interaction effects

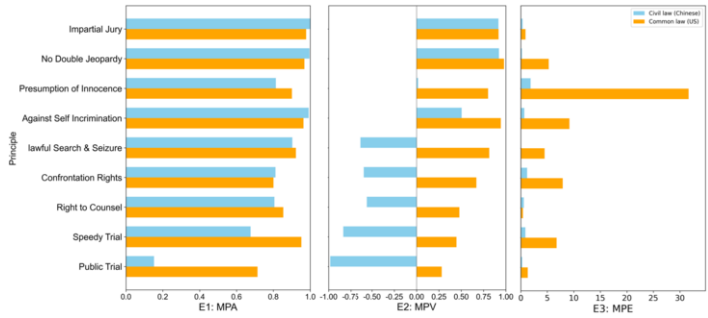



**Figure 1.** Mean of procedural principles by legal traditions

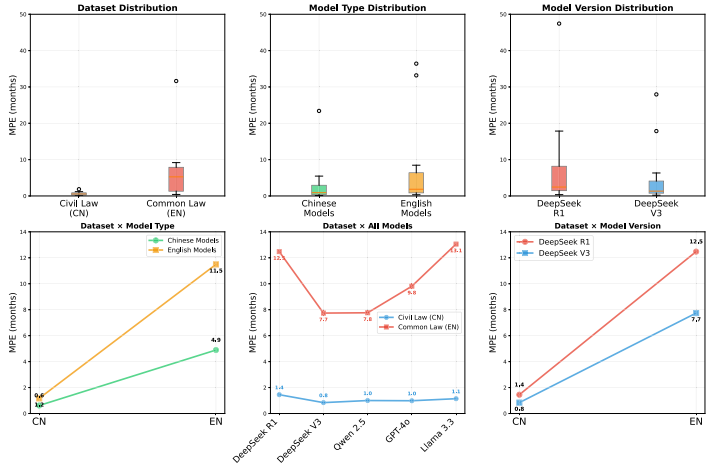



**Figure 2.** Two-Way ANOVA Analysis of Model and Dataset Interaction on E3

ence between procedural and substantive justice. Models tested on U.S. data favor procedural fairness; those tested on Chinese data prioritize substantive outcomes and often overlook procedural flaws. Deepseek R1 amplifies this difference compared to Deepseek V3. Fig 1 shows core procedural fairness principles such as presumption of innocence and protection against self-incrimination are consistently upheld. E3 employs mean of the Procedure effect on the substantive sentence MPE $=$ mean $|S_{\text{compliant}} - S_{\text{violation}}|$to quantify the impact of procedural compliance on sentencing. Models tested on U.S. datasets and using R1 show higher sensitivity to procedural impact, especially when model and dataset share the same legal tradition background (Fig 2).

## References


[1] Perry WL, McInnis B, Price CC, Smith SC, Hollywood JS. Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations. RAND Corporation; 2013. Available from: https://www.jstor.org/stable/10.7249/j.ctt4cgdcz.

[2] Chalkidis I, Androutsopoulos I, Aletras N. Neural Legal Judgment Prediction in English. arXiv; 2019. ArXiv:1906.02059 [cs]. Available from: http://arxiv.org/abs/1906.02059.

[3] Zhong H, Guo Z, Tu C, Xiao C, Liu Z, Sun M. Legal Judgment Prediction via Topological Learning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics; 2018. p. 3540-9. Available from: http://aclweb.org/anthology/D18-1390.

[4] Balkir E, Kiritchenko S, Nejadgholi I, Fraser K. Challenges in Applying Explainability Methods to Improve the Fairness of NLP Models. In: Verma A, Pruksachatkun Y, Chang KW, Galstyan A, Dhamala J, Cao YT, editors. Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022). Seattle, U.S.A.: Association for Computational Linguistics; 2022. p. 80-92. Available from: https://aclanthology.org/2022.trustnlp-1.8/.

[5] Angwin J, Larson J, Mattu S, Kirchner L. Machine Bias *. In: Ethics of Data and Analytics. Auerbach Publications; 2022. Num Pages: 11.

[6] Höddinghaus, Miriam, Sondern, Dominik, Hertel, Guido. The automation of leadership functions: Would people trust decision algorithms? | Marketing Center Münster. Placeholder Journal. 2025 Oct. Available from: https://www.marketingcenter.de/en/research/publications/164381.

[7] Eigner E, Händler T. Determinants of LLM-assisted Decision-Making. arXiv; 2024. ArXiv:2402.17385 [cs]. Available from: http://arxiv.org/abs/2402.17385.

[8] Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A Survey on Bias and Fairness in Machine Learning. ACM Comput Surv. 2021 Jul;54(6):115:1-115:35. Available from: https://doi.org/10.1145/3457607.

[9] Morse L, Teodorescu MHM, Awwad Y, Kane GC. Do the Ends Justify the Means? Variation in the Distributive and Procedural Fairness of Machine Learning Algorithms. Journal of Business Ethics. 2022 Dec;181(4):1083-95. Available from: https://link.springer.com/10.1007/s10551-021-04939-5.

[10] Grgić-Hlača N, Zafar MB, Gummadi KP, Weller A. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. Proceedings of the AAAI Conference on Artificial Intelligence. 2018 Apr;32(1). Available from: https://ojs.aaai.org/index.php/AAAI/article/view/11296.

[11] Zhao Y, Wang Y, Derr T. Fairness and Explainability: Bridging the Gap towards Fair Model Explanations. Proceedings of the AAAI Conference on Artificial Intelligence. 2023 Jun;37(9):11363-71. Available from: https://ojs.aaai.org/index.php/AAAI/article/view/26344.

[12] Wang Z, Huang C, Yao X. Procedural Fairness in Machine Learning. arXiv; 2024. ArXiv:2404.01877 [cs]. Available from: http://arxiv.org/abs/2404.01877.

[13] Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A Survey of Methods for Explaining Black Box Models. ACM Comput Surv. 2018 Aug;51(5). Available from: https://doi.org/10.1145/3236009.

[14] Weidinger L, McKee KR, Everett R, Huang S, Zhu TO, Chadwick MJ, et al. Using the Veil of Ignorance to align AI systems with principles of justice. Proceedings of the National Academy of Sciences. 2023 May;120(18):e2213709120. Publisher: Proceedings of the National Academy of Sciences. Available from: https://www.pnas.org/doi/10.1073/pnas.2213709120.

[15] David R. Les grands systèmes de droit contemporains. Paris : Dalloz; 1988. Available from: http://archive.org/details/lesgrandssysteme0000davi.

[16] Tao Y, Viberg O, Baker RS, Kizilcec RF. Cultural bias and cultural alignment of large language models. PNAS Nexus. 2024 Sep;3(9):pgae346. Available from: https://doi.org/10.1093/pnasnexus/pgae346.

[17] Cao Y, Zhou L, Lee S, Cabello L, Chen M, Hershcovich D. Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study. In: Dev S, Prabhakaran V, Adelani DI, Hovy D, Benotti L, editors. Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP). Dubrovnik, Croatia: Association for Computational Linguistics; 2023. p. 53-67. Available from: https://aclanthology.org/2023.c3nlp-1.7/.

[18] Teichman D, Ritov I, Zamir E. Biases in Legal Decision Making: Comparing Prosecutors, Defense Attorneys, Law Students, and Laypersons [SSRN Scholarly Paper]. Rochester, NY: Social Science Research Network;. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4483376.

[19] Rawls J. A Theory of Justice: Original Edition. Harvard University Press; 1971. Available from: https://www.jstor.org/stable/j.ctvjf9z6v.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*

*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251617*

# Towards Interpretable Hierarchical Graph Neural Networks in Criminal Law

Thiago DAL PONT [a,c,1], Isabela SABO [b], Maite VIEIRA [b], Jomi HÜBNER [a], and Aires ROVER [b]

[a] *Automation and Systems Department, Federal University of Santa Catarina, Brazil*
[b] *Department of Law, Federal University of Santa Catarina, Brazil*
[c] *AlmaAI Research Centre, University of Bologna, Italy*

**Abstract.** This paper proposes and evaluates a method for interpreting hierarchical graph neural network (GNN) nodes created from criminal law texts. We optimise a hierarchical GNN with extra loss functions to generate coherent abstract nodes from Brazilian habeas corpus data. Four concept grounding (CG) methods (utilising semantic similarity and large language models (LLMs)) are evaluated for giving names to these nodes. Quantitative results demonstrate that optimisation improves the quality of abstractions and predictive performance. Qualitative expert evaluation favoured labels based on LLMs. This dual approach (optimisation and grounding) provides a framework towards more trustworthy and explainable legal AI systems.

**Keywords.** Criminal Law, Interpretability, Textual Graphs, Concept Grounding

## 1. Introduction

Explainability is crucial in high-stakes domains, such as law, where transparency is required for due process [1,2]. However, standard natural language processing (NLP) methods for explainable artificial intelligence (XAI), such as SHAP [3] or LIME [4], often provide feature attributions as lists of important but structurally unrelated words. This fails to capture the relational nature of legal writing [5].

Relational explanations using graph neural networks (GNNs) on text graphs (where nodes are words and edges are dependencies) offer a promising alternative by explicitly modelling text structure [6]. However, while hierarchical GNNs can learn high-level abstractions by grouping nodes, the conceptual meaning of these abstractions is often unclear, necessitating concept grounding (CG), i.e. mapping these abstractions to real-world concepts. Our work uniquely combines GNN training optimisation aiming at coherent abstractions with a comparative evaluation of CG strategies, building on recent work in legal AI transparency [7,8] and GNN interpretability [9,10]. We thus investigate: *to what extent can we map the abstract nodes of a hierarchical GNN into natural language to explain how it reaches its predictions in criminal law cases?*

[1] Corresponding Author: Thiago Dal Pont, thiagordalpont@gmail.com. Code and materials at: https://github.com/thiagordp/TextualHGNN/tree/jurix2025

Our methodology involves generating and interpreting hierarchical representations from a corpus of Brazilian Supreme Court pre-trial detention decisions, framed as a binary classification (*released* vs. *not released*), having 2018 documents: 515 released and 1503 not released (see [11]). Decisions become graph structures using dependency parsing [12], where nodes are words (initialised with pre-trained legal GloVe embeddings, see [13,14]) and edges are syntactic dependencies. We employed DiffPool [15], a hierarchical GNN, to learn multi-level graph representations by clustering words (level 0) into abstract nodes at higher levels (Figure 1; abstract nodes to be grounded: A, B and C).

We trained GNN using a composite loss function ($L_{total} = L_{predictive} + \sum \lambda_i L_i$) having a classification loss ($L_{predictive}$) with regularisation terms ($L_i$), to balance predictive performance with interpretability. Such terms were designed for structural faithfulness, enhance abstract node quality, and embedding diversity [16,17,15]. The optimisation metric was the Interpretability-Aware Performance Score (IAPS), a hybrid metric balancing Macro F1-score with abstraction quality). We then propose four Concept Grounding (CG) methods: (1) semantic similarity to the average embedding of assigned words; (2) semantic similarity to the abstract node's embedding; (3) LLM-based (LLaMa 3.3 70B); and (4) using the top assigned word.

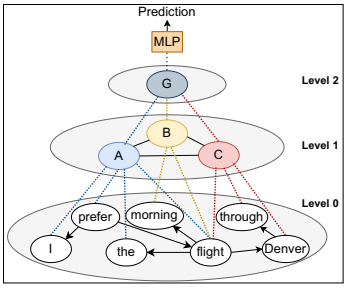



**Figure 1.** Hierarchical graph

Quantitatively, Macro F1 was measured against baselines, and IAPS assesses abstract node quality using Concept Completeness (expressiveness) and Concept Conformity (purity) [9]. For qualitative assessment, two legal experts evaluated 18 abstract nodes on semantic coherence (1-5 Likert scale) and ranked the labels from all four CG methods based on faithfulness, specificity, and plausibility.

## 2. Results, Discussion and Conclusion

Our optimised Diffpool (Macro-F1 0.821) outperformed standard GCN (0.817) baselines, while improving abstract node quality (Concept Conformity rose to 0.890, $p < 0.05$; Concept Completeness was 1.000).

Qualitatively, experts rated the GNN's abstractions as only moderately coherent (2.82/5). But, they clearly preferred CG labels from the LLM-Based and Similarity from Assigned Nodes methods, while the poor performance of using the GNN's own abstract embeddings directly for similarity search indicates they lack direct semantic grounding.

Returning to our research question, abstract nodes of a hierarchical GNN may be translated into natural language to a moderate but promising extent. Our dual approach is a viable path towards more transparent legal AI, with LLM-based and semantic search methods showing the most success. The moderate coherence of the learned abstractions and the small scale of our expert evaluation are key limitations requiring further validation. Future work will focus on integrating subgraph explanations (e.g., SubgraphX [18]), multi-level explanations [19], regression [20] and expanding expert evaluation.

**Acknowledgement:** Paper partially supported by CAPES (Coordination for the Improvement of Higher Education Personnel) and by the "CompuLaw" project, funded by the European Research Council (Grant No. 833647).

## References


[1] Bibal A, Lognoul M, de Streel A, Frénay B. Legal requirements on explainability in machine learning. Artificial Intelligence and Law. 2020 Jul;29(2):149–169.

[2] Zhang C, Meng Y. Bridging the divide: technical research and application on legal judgment prediction. Artificial Intelligence and Law. 2025 Jul.

[3] Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc.; 2017. .

[4] Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?". In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2016. p. 1135-44.

[5] Shao HL, Wang Wh, Huang Sc. Explainable AI for Real Legal Judgments: Sentencing Model and Salience Maps for Text. In: Legal Knowledge and Information Systems. IOS Press; 2024. .

[6] Wang K, Ding Y, Han SC. Graph neural networks for text classification: a survey. Artificial Intelligence Review. 2024 jul;57(8):190.

[7] Ryberg J. Artificial intelligence at sentencing: when do algorithms perform well enough to replace humans? AI and Ethics. 2024 Feb;5(2):1009–1018.

[8] Cavus M, Benli MN, Altuntas U, Sari M, Ayan H, Ugurluoglu YF. Transparent and bias-resilient AI framework for recidivism prediction using deep learning and clustering techniques in criminal justice. Applied Soft Computing. 2025;176:113160.

[9] Jürß J, Magister LC, Barbiero P, Lio P, Simidjievski N. Everybody Needs a Little HELP: Explaining Graphs via Hierarchical Concepts. In: NeurIPS 2023: New Frontiers in Graph Learning; 2023. .

[10] Yan H, Gui L, He Y. Hierarchical Interpretation of Neural Text Classification. Computational Linguistics. 2022;48(4):987–1020.

[11] Dal Pont TR, Sabo IC, Vigneaux Wilton PE, de Menezes VA, Copetti R, Zambrota L, et al. Predicting pre-trial detention outcomes in the Brazilian Supreme Court. Journal of Data Intelligence. 2022 Feb;3(1):116–130.

[12] Osman AH, Barukub OM. Graph-Based Text Representation and Matching: A Review of the State of the Art and Future Challenges. IEEE Access. 2020;8:87562–87583.

[13] Dal Pont TR, Sabo IC, Hübner JF, Rover AJ. Impact of Text Specificity and Size on Word Embeddings Performance: An Empirical Evaluation in Brazilian Legal Domain. In: Cerri R, Prati RC, editors. Intelligent Systems. Cham: Springer International Publishing; 2020. p. 521-35.

[14] Pennington J, Socher R, Manning C. GloVe: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics; 2014. p. 1532-43.

[15] Ying R, You J, Morris C, Ren X, Hamilton WL, Leskovec J. Hierarchical graph representation learning with differentiable pooling. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. NIPS'18. Red Hook, NY, USA: Curran Associates Inc.; 2018. p. 4805–4815.

[16] Godwin J, Schaarschmidt M, Gaunt A, Sanchez-Gonzalez A, Rubanova Y, Veličković P, et al. Simple GNN Regularisation for 3D Molecular Property Prediction and Beyond. In: Proceedings of the 10th International Conference on Learning Representations (ICLR); 2022. .

[17] Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005 nov;67(5):768-8.

[18] Yuan H, Yu H, Wang J, Li K, Ji S. On Explainability of Graph Neural Networks via Subgraph Explorations. In: Meila M, Zhang T, editors. Proceedings of the 38th International Conference on Machine Learning. vol. 139 of Proceedings of Machine Learning Research. PMLR; 2021. p. 12241-52.

[19] Gao J, Wang X, Wang Y, Xie X. Explainable Recommendation through Attentive Multi-View Learning. Proceedings of the AAAI Conference on Artificial Intelligence. 2019 Jul;33(01):3622–3629.

[20] Dal Pont TR, Sabo IC, Hübner JF, Rover AJ. Regression applied to legal judgments to predict compensation for immaterial damage. PeerJ Computer Science. 2023;9:e1225.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251618*

# A Tool for Handling Unexpected Exceptions in Legal Reasoning

Thomas ECOBICHON [a], Mathieu CARPENTIER [b], Sylvie DOUTRE [a], Jean-Guy MAILLY [a]
[a] *IRIT, Université Toulouse Capitole,* `jean-guy.mailly@irit.fr`
[b] *IMH, Toulouse School of Law, Université Toulouse Capitole*

**Abstract.** We have developed a system, based on prioritized default logic, for representing legal rules and their exceptions. Our system can help the law practitioner's reasoning in presence of scenarios that induce an exception to a rule. When there is no known exception to the rule $r_1$ for a scenario, the system can help the user to adapt the exception $r_2'$ to a general rule $r_2$.

**Keywords.** Default Logic, Handling exceptions, System description

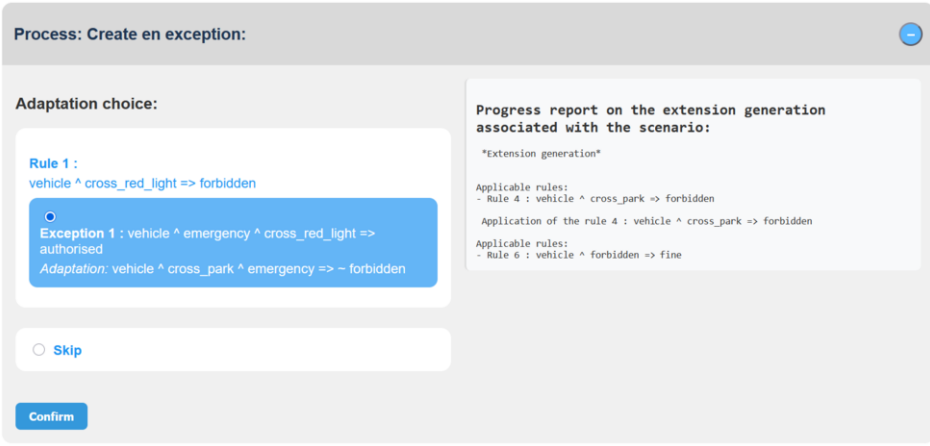
Legal reasoning is highly sensitive to *exceptions*, *i.e.* situations where a general rule would imply a given decision, but a more specific context leads to a different decision. This kind of reasoning can be modelled with defeasible reasoning approaches [1], in particular (prioritized) default logic [2,3,4], where the exception to a general rule has a higher priority than the general rule. Now the question is: what happens if the exception has not been anticipated? Consider this simple example: a rule $r_1$ forbids all vehicles to drive through the park (for instance, implying a fine for the driver), and the driver of an ambulance has decided to cross the park because traffic jam was preventing her to reach the hospital fast enough to save her patient. A strict application of the rule says that this action was forbidden (and thus the driver should receive a fine). But ethics would say that saving the life of the patient was an emergency situation, which should lead to an exception. Our goal is thus to propose to the law practitioner (*e.g.* a judge) new exceptions based on existing ones: imagine that another rule $r_2$ forbids all vehicles to cross at a red light, with the exception $r_2'$ that some vehicles (police cars, fire trucks, ambulance) can cross the red light in emergency situation. Then we can adapt this exception to the case of the park, and propose to the judge to use this new exception (and thus the ambulance driver would not receive a fine).

Formally, we assume simple *default rules* [2] $\mathscr{R} = \{r_1, \ldots, r_n\}$ where each $r_i$ is of the form $(p_1 \wedge \ldots \wedge p_k) \rightsquigarrow c$. All $p_i$'s are propositional literals called the *premises* of the rule ($\mathscr{P}(r_i)$), and $c$ is a literal called the *conclusion* of the rule ($\mathscr{C}(r_i)$). We assume a consistent set of premises, *i.e.* $\mathscr{P}(r_i) \nvdash \bot$. Given a rule base $\mathscr{R}$ and two rules $r_i, r_j \in \mathscr{R}$, $r_i$ is an *exception* to $r_j$ if $\mathscr{P}(r_j) \subset \mathscr{P}(r_i)$ and $\mathscr{C}(r_i) \wedge \mathscr{C}(r_j) \vdash \bot$. Rules are ranked with a priority relation $\succ$ [3], *i.e.* if both $r_i$ and $r_j$ could be applicable, and $r_i \succ r_j$, then only $r_i$ is actually applicable. General knowledge and the current scenario can be encoded in a propositional knowledge base $\mathscr{W}$.

To select relevant rules for adapting their exceptions, we use the Hamming distance between premises of rules: $d_{\mathscr{R}}(r_1, r_2) = d_H(\mathscr{P}(r_1), \mathscr{P}(r_2))$ if $\mathscr{C}(r_1) = \mathscr{C}(r_2)$, and $\delta$, a

fixed constant larger than the number of propositional atoms in the vocabulary, otherwise. Various policies can be used, selecting any $r_2$ with a minimal distance to $r_1$, or also taking into account a threshold. The choice of policy can be left to the user (law practitioner). Then, to adapt the exception $r_2'$ of $r_2$, to $r_1$, given the simple nature of our rules, we can use the following adaptation method: $\mathscr{A}dapt(r_1, r_2, r_2') = \mathscr{P}(r_1) \wedge (\mathscr{P}(r_2') \setminus \mathscr{P}(r_2)) \rightsquigarrow \mathscr{C}(r_2')$.

The system that we have developed consists in a Web interface that lets the user describe a scenario in natural language. An LLM parses this scenario into a set of facts which is added to the current knowledge base. Then, the rules in the current rule base which are applicable are shown to the user, which can iteratively select rules in order to obtain conclusions (what is called an *extension*) from the default theory [2,3]. At each step, the user can also choose to create a new exception using the formal above-mentioned method; see the Figure below. The updated rule base including new exceptions can be saved for future usage. The source code of our system is available online,[1] as well as a Web interface[2] and a demonstration video.[3] The current prototype contains the logical encoding of our "ambulance" toy example, and some extracts of the French penal code.




We consider several interesting future works. The first one consists in encoding in logic more legal texts. This may require a richer logical language than the restricted default logic used here; *e.g.*, with more complex rules than normal defaults made of a conjunction of literals as premises, and a single literal as conclusion, or using rule-based deontic logic. We also want to study other approaches for selecting the best rule for creating new exceptions, and for adapting their existing exceptions into new ones. Finally, using an LLM for rephrasing the new exception into natural language would be a nice improvement to the current web interface.

[1] https://github.com/Ecobichon-Thomas/Exceptions-Legal-Reasoning.git
[2] https://exceptions-legal-reasoning.onrender.com/
[3] https://github.com/Ecobichon-Thomas/Exceptions-Legal-Reasoning/blob/main/presentation_video.mp4

## Acknowledgements

This work has been funded by the French National Research Agency (grant AIDAL, ANR-22-CPJ1-0061-01) and the International Research for Society Institute (IRSI, grant Elan+). The authors thank Karla Salas-Jimenez for her work on the deployment of the Web platform.

## References

[1] Calegari R, Contissa G, Lagioia F, Omicini A, Sartor G. Defeasible Systems in Legal Reasoning: A Comparative Assessment. In: Araszkiewicz M, Rodríguez-Doncel V, editors. Legal Knowledge and Information Systems - JURIX 2019: The Thirty-second Annual Conference, Madrid, Spain, December 11-13, 2019. vol. 322 of Frontiers in Artificial Intelligence and Applications. IOS Press; 2019. p. 169-74. Available from: https://doi.org/10.3233/FAIA190320.

[2] Reiter R. A Logic for Default Reasoning. Artificial Intelligence. 1980;13(1-2):81-132. Available from: https://doi.org/10.1016/0004-3702(80)90014-4.

[3] Brewka G. Reasoning about Priorities in Default Logic. In: Hayes-Roth B, Korf RE, editors. Proceedings of the 12th National Conference on Artificial Intelligence, Seattle, WA, USA, July 31 - August 4, 1994, Volume 2. AAAI Press / The MIT Press; 1994. p. 940-5. Available from: http://www.aaai.org/Library/AAAI/1994/aaai94-144.php.

[4] Lawsky SB. A Logic for Statutes. Florida Tax Review. 2017;21(1). Available from: https://scholarship.law.ufl.edu/ftr/vol21/iss1/2.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*

*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251619*

# Normative Risk Assessment in the AI Act

Beatrice FERRIGNO [a,1], Marco BILLI [a], Yasaman YOUSEFI [a], and Antonino ROTOLO [a]

[a] *Alma AI - University of Bologna, Italy*

ORCiD ID: Beatrice Ferrigno https://orcid.org/0009-0007-9197-2970, Marco Billi https://orcid.org/0000-0002-6807-073X, Yasaman Yousefi https://orcid.org/0000-0003-1483-2978, Antonino Rotolo https://orcid.org/0000-0001-5265-0660

**Abstract.** This work introduces the formal framework **DLRisk**, based on Defeasible Deontic Logic, for assessing how fundamental rights are balanced under the EU AI Act when new systems are developed. This approach enables qualitative, scenario-aware reasoning by ordering rights, capturing demotions triggered by factual conditions, and detecting collisions between rights and deployment contexts. The result is an explainable assessment of trade-offs that supports compliance analysis and risk mitigation.

**Keywords.** Risk assessment, Defeasible deontic logic, Fundamental rights

## 1. Introduction

Risk assessment is the core idea behind the European AI Act (AIA), the normative framework that has been designed to regulate the introduction of AI systems on the basis of their risk profiles. However, a key methodological issue remains: how can we evaluate and mitigate the impact of technical hazards on fundamental rights?

This question belongs to a long tradition on balancing conflicting interests, where technical progress may represent an issue for rights' protection. In the context of the AI Act risk mitigation, this work tries to propose a methodology to investigate how specific features of an AI system may demote certain rights or generate collision between them. This project aims to be a balancing test through which legal experts can evaluate whether the benefits of an AI introduction overweighs or not detriments.

Following [1,2] we argue for an analysis based on risk scenarios involving interactions among multiple risk factors, rather than a scope-oriented characterization. Building on [3], we thus propose a way to formalize a balancing evaluation with formal method, to avoid using opaque approaches to perform an interpretable analysis, that must support a human-in-the-loop approach.

[1] Corresponding Author

## 2. The framework

For this work, we propose a framework based on Defeasible Deontic Logic, **DLRisk**. This framework is able to model reasoning patterns with a large spectrum of legal concepts and issues [4], and with our operators is possible to organize rights hierarchically, from the most essential to protect to the lower priorities, while linking their status to concrete deployment scenarios.

*Language* Let PROP be propositional atoms and $\text{Lit} = \text{PROP} \cup \{\neg p \mid p \in \text{PROP}\}$ the descriptive literals for scenarios (e.g., $S_i$). Let $\mathbf{Rights} = \{\mathbf{r^1}, \mathbf{r^2}, \dots\}$ be the set of legal rights (AIA). Following [3,5], rights are prioritized via $n$-ary operators $\otimes^n$ ($n > 0$). Let **Pref** be the set of all $\otimes$-expressions. The reading of $\mathbf{r^1} \otimes \mathbf{r^2}$ is: prefer $\mathbf{r^1}$; if demoted, prefer $\mathbf{r^2}$. We annotate positions by subscripts, e.g., $\mathbf{r^j_1} \otimes \mathbf{r^k_2} \otimes \cdots \otimes \mathbf{r^i_n}$.

*Operators* We introduce a binary operator Promotes with degree: $\mathsf{Promotes} : \text{Lit}^2 \times \{-2,-1,0,1,2\} \times \mathbf{Rights} \to$ expressions. $\mathsf{Promotes}(X, n, \mathbf{r^j})$ means that $X$ promotes/demotes $\mathbf{r^j}$ with degree $n$. For a more descriptive language, we write $\mathsf{Promotes}(Y, \mathbf{r^k})$ for $n > 0$ and $\mathsf{Demotes}(Y, \mathbf{r^k})$ for $n < 0$. Finally, if $X \subseteq \text{Lit}$, we write $\mathsf{Collide}(X, \mathbf{r^x}, \mathbf{r^y})$ to denote that the rights $\mathbf{r^x}, \mathbf{r^y}$ normatively collide with respect to $X$ [6].

For any literal $l$, let $\tilde{l}$ be its complement. A set $Y$ is *consistent* iff it contains no complementary pair. Let **PrDe** and **Coll** be the sets of (possibly complemented) promotion/demotion and collision literals, respectively. Lastly, we define *rules*. Let Lab be a set of labels to represent names of rules: a rule $r$ has the form $r\colon A(r) \Rightarrow C(r)$, where: (i) $r \in \text{Lab}$ is the unique name of the rule, (ii) $A(r) \subseteq \text{Lit} \cup \mathbf{PrDe}$ is $r$'s (set of) antecedents, (iii) $C(r) = \phi \in \text{Lit} \cup \mathbf{PrDe} \cup \mathbf{Pref}$ is its conclusion.

## 3. Heuristics to Minimize Legal Risk

We elaborate on [3]'s metrics, to propose a weighted quantitative approach to evaluate the overall impact on fundamental rights. For any deployment system $\mathscr{D}_\sigma$ we define the *degree of right impact* $\mathbf{Degree}(\mathscr{D}_\sigma) = \Xi - \Delta$, where $\Xi$ is the sum of the defeasibly derived rights, plus an additional weight if $D_{\mathscr{D}_\sigma} \vdash \mathsf{Promotes}(Y, w, \mathbf{r}_{\langle x,y\rangle})$ and $0 < w \leq 2$. $\Delta$ applies the same calculation to the demoted rights, so where $-2 < w \leq 0$, yielding an estimate of the costs incurred by adopting AI in the given scenarios. This quantitative summary helps policymakers assess impact at a glance (e.g., a negative value signals an excessive rights burden).

## 4. Conclusions

We presented a transparent, qualitative method for evaluating conflicts between rights protection and AI deployment. Although we have not yet validated DLRisk on real-world datasets, it offers a principled basis for balancing under the AIA and for guiding future compliance and mitigation strategies. A limitation of this work concerns the rights, which cannot be exhaustively captured by formal operators. Accordingly, the framework assumes a *human-in-the-loop* who supplies domain knowledge, weighs trade-offs, and can override or refine symbolic outputs [7,8].

## References

[1] Novelli C, Casolari F, Rotolo A, Taddeo M, Floridi L. Taking AI risks seriously: a new assessment model for the AI Act. AI & Society. 2023. Available from: https://doi.org/10.1007/s00146-023-01723-z.
[2] Novelli C, Casolari F, Rotolo A, Taddeo M, Floridi L. AI Risk Assessment: A Scenario-Based, Proportional Methodology for the AI Act. Digital Society. 2024;3:1-29.
[3] Rotolo A, Ferrigno B, Godinez JMAG, Novelli C, Sartor G. Foundations for Risk Assessment of AI in Protecting Fundamental Rights; 2025. Available from: https://arxiv.org/abs/2507.18290.
[4] Governatori G, Rotolo A, Sartor G. Logic and the Law: Philosophical Foundations, Deontics, and Defeasible Reasoning. In: Handbook of Deontic Logic and Normative Reasoning. vol. 2. London: College Publications; 2021. p. 655-760.
[5] Calardo E, Governatori G, Rotolo A. Sequence Semantics for Modelling Reason-based Preferences. Fundam Informaticae. 2018;158(1-3):217-38. Available from: https://doi.org/10.3233/FI-2018-1647.
[6] Alexy R. Theorie der Grundrechte. Frankfurt am Main: Suhrkamp; 1994.
[7] Koulu R. Proceduralizing control and discretion: Human oversight in artificial intelligence policy. Maastricht Journal of European and Comparative Law. 2020;27(6):720-35.
[8] Binns R. Human Judgment in algorithmic loops: Individual justice and automated decision-making. Regulation & governance. 2022;16(1):197-211.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251620*

# An Argumentative Explanation Framework for Generalized Reason Model with Inconsistent Precedents

Wachara FUNGWACHARAKORN [a] Gauvain BOURGNE [b] Ken SATOH [a]
[a] *Center for Juris-Informatics, ROIS-DS, Tokyo, Japan*
[b] *LIP6, Sorbonne University, CNRS, Paris, France*

**Abstract.** Precedential constraint is one foundation of case-based reasoning in AI and Law. It generally assumes that the underlying set of precedents must be consistent. To relax this assumption, a generalized notion of the reason model has been introduced. While several argumentative explanation approaches exist for reasoning with precedents based on the traditional consistent reason model, there has been no corresponding argumentative explanation method developed for this generalized reasoning framework accommodating inconsistent precedents. To address this question, this paper examines an extension of the derivation state argumentation framework (DSA-framework) to explain the reasoning according to the generalized notion of the reason model.

**Keywords.** argumentation, explanation, precedential constraint, factors, inconsistent case-base

## 1. Background: The Generalized Reason Model

The **reason model** assumes that a dispute has two *sides*: *plaintiff* $\pi$, and *defendant* $\delta$. It also assumes a factor domain $\mathscr{F}$, partitioned into pro-plaintiff factors $\mathscr{F}^{\pi}$ and pro-defendant factors $\mathscr{F}^{\delta}$. A *case* is represented as a tuple $\langle X, r, s\rangle$ where $X$ is a fact situation; $s$ is the *winning* side; and $r = U \rightarrow s$ is a *rule*, where $U$ is the reason set (subset of pro-$s$ factors in $X$). For any side $s \in \{\pi, \delta\}$ and any fact situation $X$, the opposite side of $s$ is denoted as $\bar{s}$ (i.e., $\pi = \bar{\delta}$ and $\delta = \bar{\pi}$) and the set of all pro-$s$ factors in $X$ is denoted as $X^s$ (i.e., $X^s = X \cap \mathscr{F}^s$).

A case $c$ induces *a fortiori* priority $<_c$ over reason sets. Specifically, for $U \subseteq \mathscr{F}^{\bar{s}}$ and $V \subseteq \mathscr{F}^s$, $U <_c V$ if $U \subseteq X^{\bar{s}}$ and $premise(r) \subseteq V$. A case base (a finite set of cases) $\Gamma$ induces a priority $<_\Gamma$ over reason sets $U$ and $V$ such that $U <_\Gamma V$ if there is a case $c$ in $\Gamma$ such that $U <_c V$. An *inconsistency* in $\Gamma$, denoted as $U \perp_\Gamma V$, occurs if the case base induces priority such that $U <_\Gamma V$ and $V <_\Gamma U$ for some reason set $U$ and $V$. $inc(\Gamma)$ denotes all inconsistencies in $\Gamma$. A case base is consistent if $inc(\Gamma) = \emptyset$, and inconsistent otherwise.

The generalized reason model [1], crucial for handling inconsistency, defines that a side $s$ is *permitted* $\Gamma \mid\!\sim P_X(s)$ if there exists a permitted rule $r$ such that adding the resulting case does not create new inconsistencies: $inc(\Gamma \cup \{\langle X, r, s\rangle\}) \subseteq inc(\Gamma)$, and a side $s$ is *obligated* $\Gamma \mid\!\sim O_X(s)$ if $s$ is the only side permitted.

## 2. DSA-Framework for Generalized Reason Model

The proposed framework extends the original DSA-framework [2] as follows:

- **Conclusiveness and Derivation State**: A case base $\Gamma$ is *conclusive* with respect to $X$ ifthere exists $s \in \{\pi, \delta\}$ such that $X^{\bar{s}} <_\Gamma X^s$ and $X^s \not<_\Gamma X^{\bar{s}}$ and $s$ is called a *derivation state* for $X$ with respect to $\Gamma$.
- **Maximal Inconsistent Sub-base**: Given a case base $\Gamma$, A maximal conclusive sub-base is a $\supseteq$-maximal subset of $\Gamma$ that remains conclusive w.r.t. $X$, denoted $max(\Gamma, X)$.
- **Derivation state argument** (DS-argument) is a tuple $(\chi, \gamma, s) \in 2^X \times 2^\Gamma \times \{\pi, \delta\}$ such that (a) $\chi \subseteq X$; (b) $\gamma \in max(\Gamma, \chi)$; and (c) $s$ is a derivation state for $\chi$ with respect to $\gamma$.
- **DSA-framework** is an abstract argumentation framework $(\mathscr{D}_{X,\Gamma}, \mathscr{R})$ where $\mathscr{D}_{X,\Gamma}$ is the set of all DS-arguments. The attack relation $R$ is defined such that argument $(\chi, \gamma, s)$ attacks $(\chi', \gamma', s')$ if and only if
  - (change derivation state) $s \neq s'$, and
  - (gain more knowledge) $\chi' \subsetneq \chi$, and
  - (concise attack) no $(\chi'', \gamma'', s)$ in $\mathscr{D}_{X,\Gamma}$ with $\chi' \subsetneq \chi'' \subsetneq \chi$.
- **Argumentative Explanation** for why $s$ is obligated is defined as any admissible dispute tree for a non-attacking DS-argument with derivation state $s$. The set of all explanations is denoted as as $\mathscr{E}_{\Gamma,X}(s)$.

**Theorem**: $\Gamma \mathrel{|\!\sim} O_X(s)$ if and only if $\mathscr{E}_{\Gamma,X}(s) \neq \emptyset$ and $\mathscr{E}_{\Gamma,X}(\bar{s}) = \emptyset$.

The theorem means that a side is obligated if and only if there exists an explanation for the obligated side and no explanation exists for the opposite side.

**Future Work**: Several challenges mentioned remain unanswered but we can foresee some potentials. For example, this paper has not investigated an explanation for two-side permissions (i.e., $\Gamma \mathrel{|\!\sim} P_x(s)$ and $\Gamma \mathrel{|\!\sim} P_x(\bar{s})$) but it may be possible by introducing parallel explanations (i.e., $\mathscr{E}_{\Gamma,X}(s) \neq \emptyset$ and $\mathscr{E}_{\Gamma,X}(\bar{s}) \neq \emptyset$) or rebutting attacks (two-way attacks) between arguments with the same situational knowledge but different derivation states.

**Extended Version**: https://arxiv.org/abs/2510.19263

**Acknowledgements**: This work was supported by the "R&D Hub Aimed at Ensuring Transparency and Reliability of Generative AI Models" project of the Ministry of Education, Culture, Sports, Science and Technology and JSPS KAKENHI Grant Numbers, JP22H00543, JP25H00522, JP25H01112 and JP25H01152.

## References

[1] Canavotto I. Precedential constraint derived from inconsistent case bases. In: Legal Knowledge and Information Systems. IOS Press; 2022. p. 23-32. doi:10.3233/FAIA220445.

[2] Fungwacharakorn W, Tsushima K, Hosobe H, Takeda H, Satoh K. An Argumentative Approach for Explaining Preemption in Soft-Constraint Based Norms. In: Proceedings of AICOM track of the International Workshop on AI Value Engineering and AI Compliance Mechanisms (VECOMP 2024); 2024. p. 7-12. doi:10.48550/arXiv.2409.04065.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251621*

# Uncovering AI Governance Themes in EU Policies Using BERTopic and Thematic Analysis

Delaram GOLPAYEGANI [a,1], Marta LASEK-MARKEY [a,b], Arjumand YOUNUS [c], Aphra KERR [a,c], and Dave LEWIS [a,b]

[a] *ADAPT Centre*
[b] *Trinity College Dublin, Dublin, Ireland*
[c] *University College Dublin, Dublin, Ireland*

ORCiD ID: Delaram Golpayegani https://orcid.org/0000-0002-1208-186X, Marta Lasek-Markey https://orcid.org/0000-0002-0183-3982, Arjumand Younus https://orcid.org/0000-0001-7748-2050 , Aphra Kerr https://orcid.org/0000-0001-5445-7805, Dave Lewis https://orcid.org/0000-0002-3503-4644

**Abstract.** The upsurge of policies and guidelines that aim to ensure Artificial Intelligence (AI) systems are safe and trustworthy has led to a fragmented landscape of AI governance. The European Union (EU) is a key actor in the development of such policies and guidelines. Its High-Level Expert Group (HLEG) issued an influential set of guidelines for trustworthy AI, followed in 2024 by the adoption of the EU AI Act. While the EU policies and guidelines are expected to be aligned, they may differ in their scope, areas of emphasis, degrees of normativity, and priorities in relation to AI. To gain a broad understanding of AI governance from the EU perspective, we leverage qualitative thematic analysis approaches to uncover prevalent themes in key EU documents, including the AI Act and the HLEG Ethics Guidelines. We further employ quantitative topic modelling approaches, specifically through the use of the BERTopic model, to enhance the results and increase the document sample to include EU AI policy documents published post-2018. We present a novel perspective on EU policies, tracking the evolution of its approach to addressing AI governance.

**Keywords.** AI governance, EU AI policy, topic modelling, BERTopic, thematic analysis, EU AI Act

## 1. Introduction

The rapid advancement of Artificial Intelligence (AI) technologies, coupled with an increasing number of AI-related incidents, has positioned AI as a key topic in global policymaking discourse. The European Union (EU) has become a leader in this landscape, particularly after the enactment of its AI regulation, the EU AI Act [1], in 2024. However,

[1] Corresponding Author:Delaram Golpayegani, golpayes@tcd.ie.

multiple documents were issued both before and after the adoption of the AI Act that have shaped the discourse in the EU. These range from the EU-High Level Expert Group (HLEG)'s "Ethics Guidelines for Trustworthy AI" (2019) [2], to the guidelines issued by the European Commission in 2025 on prohibited AI practices (C(2025) 5052 final) [3]. While EU AI policies are expected to be aligned in pursuit of a common vision of trustworthy AI, they adopt distinct approaches towards AI governance, which is manifested in their focus area (e.g. AI innovation, risk management, sector-specific requirements), their degree of normativity (binding, non-binding), enforcement mechanisms, and technical and organisational measures they prescribe. There is also a temporal dimension to consider, since these policies were developed at different times and therefore may not address recent advances in AI capabilities.

In this paper, we aim to uncover AI governance concepts within the EU AI policy landscape through identification of recurring topics and themes, both shared and divergent. The research question this paper explores is: *What are the prevalent themes in the EU AI policies and guidelines post-2018 concerning AI governance?*

One way of addressing the research question is through the use of *qualitative thematic analysis* approaches [4] from social sciences. Since thematic analysis is conducted manually by domain experts, it can yield nuanced and high-quality insights, however it is inherently subjective, time-intensive, and difficult to scale. To mitigate these limitations, we adopt *unsupervised topic modelling* techniques from natural language processing (NLP), which enable identification of frequent topics in large text corpora, in combination with thematic analysis to interpret, validate, and provide context to the discovered topics.

In the remainder of this paper, we first review the body of work on the analysis of AI policies in section 2. Then, in section 3 we discuss our methodology that combines topic modelling and thematic analysis approaches. section 4 provides the identified AI topics and themes. Finally, section 5 concludes the paper by discussing our plan for future work.

## 2. Related Work

There is a body of work that analyses the content of AI policies and guidelines to determine the degree of consensus (or divergence) as well as the gaps in AI governance and ethical AI principles within the growing body of AI policies. These studies apply qualitative thematic analysis, quantitative content analysis (often using topic modelling algorithms such as LDA and BERT), or a combination of both. In the following, we review some of the related work on AI policies that utilised these approaches.

### *2.1. Automated Analysis of AI Policies via Topic Modelling*

A pioneering work [5] within the realm of NLP-based AI policies' analysis uses topical distributions from LDA to perform clustering over 12 national AI strategies, and use the cluster information to determine the strategic priorities of various governments in the AI race. It is important to note that this analysis was performed over some of the earliest available documents covering strategic directions laid by national governments with respect to AI; and hence there is also more coverage on leadership in AI deployment for innovation. Another work built on LDA-based analysis and visualisations mines China's

data governance policies to discover evolution paths within these [6]. One of the uncovered themes is AI-based decision-making for data governance. Similarly, Wang et al. [7] pursue via structured topic models a comparative analysis of AI policies from China, EU and US uncovering significant differences in priority areas within the AI policymaking landscape.

### 2.2. *Combination of Automated and Manual Analysis of AI policies*

Roche et al. [8] used a combination of automated data analytics tools and manual content analysis on a corpus of AI policies and governance frameworks to investigate the inclusivity of ethical AI approaches. Specifically, inclusivity here focuses on the contribution of voices from the Global South within the AI ethics discourse. Schiff et al. [9] analysed the degree of global consensus on ethical AI principles by manual coding of AI documents published by public sector, private organisations, and NGOs, followed by a quantitative analysis of the coding. Works by Saheb and Saheb [10,11] perform a mixed method analysis of AI ethics guidelines issued by national governments to identify the most pressing concerns with respect to the ethical aspects of AI. Specifically, their methodology involves the application of topic modelling[2] followed by manual extraction of themes from within the documents. A very recent work by Kajava et al. [12] attempts to track the discourse surrounding increased risks in AI landscape following the launch of ChatGPT. Their method comprises the use of guided BERTopic in conjunction with qualitative content analysis; and this was done to ensure scalability and trust in the analysis of the discourse. Similarly, Suter et al. [13] track AI discussions in various parliamentary debates through a combination of BERTopic and manual classification into interpretable themes by two independent coders. They specifically analyse debates from US Congress, EU Parliament, Parliament of Singapore, and Swiss Federal Assembly and this helps them uncover distinct national priorities in a post-LLM world. Essentially, the findings of both works [12,13] reveal a greater focus on increasing literacy and ethical awareness within the adoption of AI policies in a world that has been shaken by software such as ChatGPT.

A key difference between our analysis and the existing ones lies in the way we track the progress made by a single entity. i.e., the European Union, which has emerged as a leader in AI governance and regulatory landscape. On a superficial level, this may seem similar to the analysis by Kajava et al. [12]. Still, their primary focus is on tracking changes in risk discourse within EU policy circles following the launch of ChatGPT. Conversely, ours seeks to understand the discourse as a whole, tracking overall continuities and shifts over time. This is especially important as implementation of the AI Act is dependent on the development of guidelines, codes of conduct, codes of practice and technical standards. The Act is also subject to changes via implementing and delegated acts as well as revisions arising from the EU Omnibus Programme that aims to minimise needless duplication of compliance and oversight effort. This presents a complex and changing body of texts that must be monitored and analysed to understand changes to the expectations and effect of the original Act. Moreover, our analysis is complemented by a comparison conducted by experts from law, computer science, and social science, resulting in refined themes that effectively capture the complexities and nuances of EU policymaking efforts regarding AI governance.

[2]They use the LDA variant embedded in the tool WordStat.

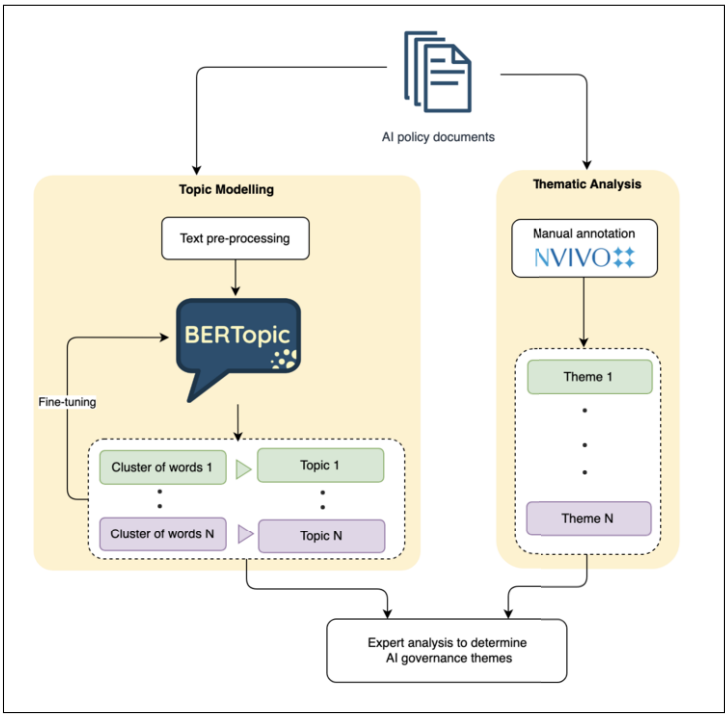



**Figure 1.** The overall methodology

## 3. Methodology

As shown in Figure 1, our methodology comprises four key steps: (1) corpus creation of EU AI policy documents, (2) manual thematic analysis, (3) unsupervised topic modelling and (4) interdisciplinary expert analysis of topics to identify themes and derive additional insights. For the purpose of this paper, we consider a *topic* as a concept identified automatically by a topic modeller from text (algorithm-driven), and a *theme* as a broad interpretable semantic concept derived through expert interpretation, either from the topics generated by a topic modeller or directly from the text itself (human-driven).

### *3.1. Corpus Creation*

Our study focuses on EU AI policy documents issued between January 2018 and August 2025, covering the period from early trustworthy AI guidelines through the recent regulation setting efforts. Our corpus includes legislation, official guidelines, and codes of practice from relevant EU institutions. Given that English is one of the official EU languages, all of the documents were published in English, which we used for our analysis. The list of documents that constitute our corpus is provided in Table 1.

### *3.2. Thematic Analysis*

Thematic analysis is a popular method of qualitative data analysis that systematically organises datasets, and helps to identify patterns of meaning commonly referred to as

**Table 1.** EU AI policy documents included in the corpus

| ID | Document | Issuer | Type | Year |
|---|---|---|---|---|
| 01 | Policy and investment recommendations for trustworthy AI [14] | EU HLEG | Recommendation | 2019 |
| 02 | Ethics guidelines for trustworthy AI [2] | EU HLEG | Guideline | 2019 |
| 03 | Assessment List for Trustworthy AI (ALTAI) [15] | EU HLEG | Guideline | 2020 |
| 04 | Sectoral considerations on policy and investment recommendations for trustworthy AI [16] | EU HLEG | Recommendation | 2020 |
| 05 | EU AI Act [1] | EU Council and Parliament | Legislation | 2024 |
| 06 | Commission Guidelines on prohibited AI practices established by the AI Act [3] | European Commission | Guideline | 2025 |
| 07 | Commission Guidelines on the scope of the obligations for general-purpose AI models established by the AI Act [17] | European Commission | Guideline | 2025 |
| 08 | EU Code of Practice for General-Purpose AI Models [18], including *transparency*, *copyright*, and *safety and security* chapters | European Commission | Code of practice | 2025 |

themes [19]. First described in the 1970s by Houlton – albeit the term itself had been in use even earlier – it became more prominent in the late 1990s with researchers like Boyatzis and Hayes [4]. In recent times, thematic analysis has been understood as an umbrella term for different approaches. While popular in qualitative interview data analysis, computer-assisted expert thematic analysis of legal texts, such as legislation or policy documents, appears to be less commonly employed.

We performed qualitative analysis of a subset of the corpus, including the EU AI Act and the HLEG "Ethics Guidelines for Trustworthy AI" with the use of NVivo software. As explained by Dhakal, NVivo is a CAQDAS programme that assists, rather than replaces a human researcher [20]. NVivo analysis is, thus, expert-led and in this project it was performed by a legal researcher. The documents were imported into NVivo and thematic coding was performed by selecting a section of text and then tagging it with a node corresponding to a theme. Rather than being preassigned ahead of time, the nodes were created *in situ*, as was deemed necessary by the human researcher. As a result, the EU AI Act analysis resulted in 59 themes, while the HLEG Guidelines resulted in 52 themes, with multiple themes overlapping across both documents.

As mentioned above, in this project thematic analysis was conducted on a subset of documents in parallel to topic modelling. The outputs of two methods were then compared, with the results of the manual thematic analysis serving to validate the topic modelling approach.

### *3.3. Topic Modelling*

Topic modelling refers to identification of prevalent topics in corpora of text using unsupervised machine learning techniques and is primarily used for uncovering topics within large sets of documents. However, this work only deals with limited number of documents that should be processed separately. Therefore, to enable applying a topic modelling algorithm on a single document, we treated each document as a set of smaller doc-

uments. For this, using the `PyMuPDF` library in Python, we extracted the text from the PDF file and then split the text into smaller parts, in this case sentences, and stored them in a CSV file.

There exist a variety of unsupervised topic models (for evolution of such models see [21]). Among the existing topic models, in consultation with NLP experts, we initially considered Latent Dirichlet Allocation (LDA) [22] and BerTopic [23]. We applied both topic modelling algorithms to a sample of policy documents (the EU AI Act, HLEG trustworthy AI guidelines, and UNESCO's guideline on the ethics of AI [24][3]) and compared their outputs against the results of manual thematic analysis. Our findings indicate that BERTopic outperforms LDA in the context of AI policy texts, offering higher accuracy and better interpretability. This is aligned with the findings of Kajava et al. [12], which compared the interpretability of these two topic modellers for EU-related documents and news articles on AI.

Before application of `BERTopic`, we employed common NLP approaches to apply minimal preprocessing using the `nltk` library for: (1) lowercasing , (2) removing non-alphabetic tokens (e.g. punctuation marks), (3) removal of tokens that consist solely of numbers, and (4) exclusion of tokens shorter than two characters. We did *not* remove stop words or perform lemmatisation prior to embedding, given that BERTopic uses transformer-based embeddings, which require the text to be persevered for better accuracy. We remove the stop words, after generating embeddings, using the `CountVectorizer` when initialising BERTopic. We fine-tuned `BERTopic` model in a iterative manner to ensure that the model identifies at least 3 topic categories for each document. The snippets used for preprocessing and topic modelling are available on GitHub under the MIT licence at https://github.com/DelaramGlp/forsee_topicmodelling.

### *3.4. Expert Analysis*

The expert analysis was conducted in two phases, each serving a different purpose:

1. **Validation of topic model**: This phase aimed to identify the degree of topic modelling accuracy by comparing its outputs with the results of manual thematic analysis. An analysis of the HLEG trustworthy AI guidelines and the EU AI Act as representative case studies was conducted by the authors, who are experts in law, computer science and social sciences. The results serve as a basis for applying automated analysis to the remaining documents in the corpus.
2. **Topic interpretation and theme discovery**: This phase involved qualitative refinement of the BERTopic results. The first and second authors of this paper, both experts in EU AI policies, independently analysed the generated topics and assigned themes to each topic category. The proposed themes were then discussed in a joint session and consolidated into their final form.

[3]Since the UNESCO recommendation is not an EU-specific policy, it is out of the scope of the work presented in this paper. However, we included it in this step as a comparative reference to ensure the generalisability of our approach across different policy contexts.

## 4. AI Governance Themes Identified in BERTopic Results

### 4.1. From Topics to Themes

Applying the BERTopic model to each document produced 227 topic clusters in total (see the breakdown of the topic clusters in Table 2). The output of the topic modelling for each document is visualised using both bar charts and word cloud digrams, and the corresponding results are stored in CSV format.

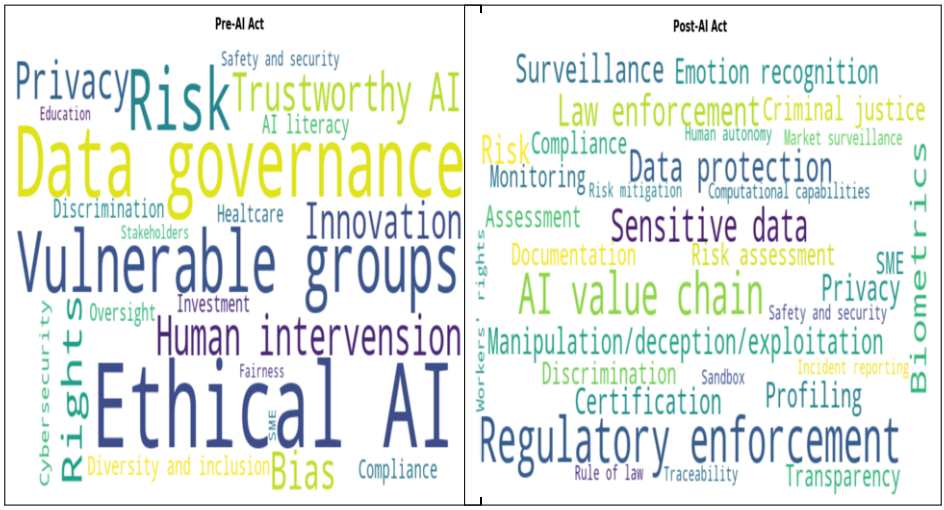
Our analysis of outputs from BERTopic was conducted manually to assign themes to each topic cluster. Despite text pre-processing and hyperparameter optimisation of BERTopic, approximately 14% of the clusters exhibited insufficient coherence to allow meaningful thematic interpretation. Additionally, while some topic clusters were coherent, they correspond to multiple themes. We adopted a minimalistic approach when assigning themes, i.e. attributing the smallest set of themes to each cluster. In cases where comprehensive representation required multiple themes, we established a constraint of assigning a maximum of 3 themes to each cluster. The analysis resulted in **100 distinct themes** in total. Table 2 shows the number of topic clusters and themes, along with the key themes identified in each EU AI policy document. Figure 2 provides an aggregated thematic overview using a word cloud visualisation, highlighting the most prevalent themes across the corpus.




**Figure 2.** Word Cloud of the identified AI governance themes from EU AI policies

### 4.2. Evolution of Themes in EU AI Policies

Our analysis shows a shift in the AI policy discourse in the EU following the adoption of the EU AI Act. As shown in Figure 3a, the pre-AI Act thematic landscape predominately focused on *ethical AI*, *data governance*, and *risks* particularly those that impact *vulnerable groups*. This emphasis aligns with both the non-binding nature of the EU HLEG documents and their mission to promote ethical AI, raise awareness of AI risks, and provide a strategic roadmap, particularly in 2019–2020, when the concept of trustworthy AI




(a) Themes pre-AI Act (b) Themes post-AI Act

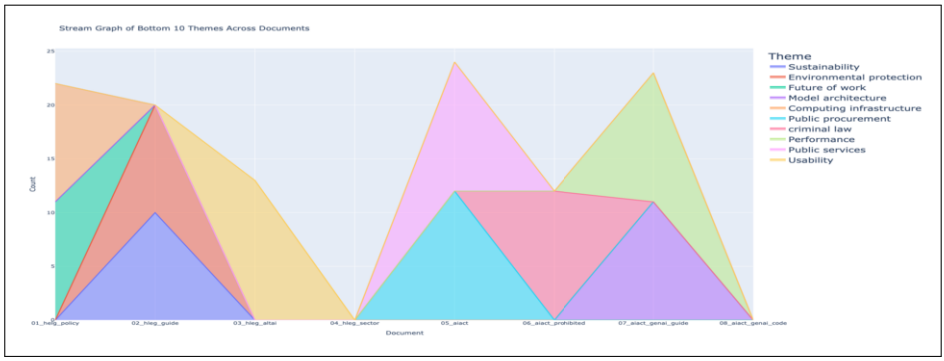
**Figure 3.** Word clouds demonstrating evolution of AI governance themes in the EU

was emerging as a popular yet still unfamiliar topic. Themes from the post-AI Act era (Figure 3b) illustrate a clear shift from *aspirational ethical AI* toward *operationalised legal AI*, with particular emphasis on *regulatory enforcement*. In addition to concrete compliance tasks such as *documentation* and *risk assessment*, the AI Act and its associated guidelines distinguish between entities that are involved in development, deployment and use of AI systems, i.e. *AI value chain*, indicating the need for involvement of multiple entities to achieve trustworthy AI. Post-AI Act themes also revealed the focus on specific AI use cases that impose significant risk. The transition in the AI policy discourse is clearly illustrated in the stream graph of top 10 themes in Figure 4b.

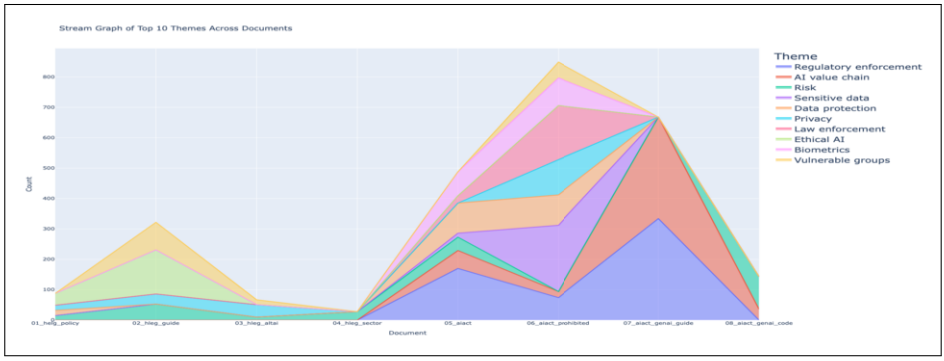
The emphasis on *risk* and *data protection* remains consistent from the pre- to post-legislation period, with the latter reflecting the influence of the EU GDPR. However, certain themes appear to be overlooked in this shift. To have a better understanding of such themes, we analysed the evolution of the bottom 10 (least mentioned) themes in Figure 4a. The stream graph shows a waning of attention to *environmental impacts* of AI as well as its broader effects on the *future of work* in the post-AI Act era.

## 5. Conclusion and Future Work

This study has demonstrated the efficacy of employing BERTopic, validated and refined through expert interdisciplinary thematic analysis, to identify the complex and evolving landscape of AI governance within the EU. The identified themes cover a wide-range, including fundamental rights, risk management, technical capabilities, regulatory enforcement, AI literacy, support for innovation, and market investment. Our work advances the state of the art in three key ways: (1) it employs a combination of topic modelling and thematic analysis enabling scalable yet human-centric approach to analyse the evolving body of AI policies, (2) it provides a holistic and temporal view of EU AI policies, and




(a) Bottom 10 themes




(b) Top 10 themes

**Figure 4.** Stream graphs demonstrating evolution of most and least prevalent topics over time

(3) it concentrates specifically on AI governance and success criteria for the responsible development and deployment of AI.

In future work, we plan to expand our analysis and investigate AI governance themes that emerged in the guidelines set forth by supranational bodies, such as UNESCO, The Organisation for Economic Co-operation and Development (OECD), and the Council of Europe. Given the importance of technical standards in the implementation of trustworthy AI guidelines as well as the EU AI Act, we also aim to identify the themes in AI standards developed by international and European standardisation bodies, namely ISO/IEC JTC 1/SC 42 and CEN/CENELEC JTC 21. Comparison of the themes identified from these two types of AI documents will assist in clarifying the interplay between technical standards and trustworthy AI guidelines and regulations. Moreover, from a methodological perspective we aim to incorporate more refined and interpretable topic models [25] leading to increased scalability across a larger set of documents. Our methodologial approach aims to optimise human-AI collaboration in policy document analysis.

**Table 2.** Overview of the topics and themes associated with each document

| ID | Document | No. of BERTopic clusters | No. of themes | Key themes |
|---|---|---|---|---|
| 01 | HLEG policy and investment recommendations [14] | 26 | 23 | AI literacy, SME, Education, Data governance, Ethical AI, Investment, Innovation |
| 02 | HLEG ethics guidelines [2] | 28 | 27 | Ethical AI, Rights, Vulnerable groups, Compliance, Risk, Fairness, Diversity and inclusion, Bias, Safety and security, Oversight, Privacy |
| 03 | HLEG ALTAI [15] | 14 | 16 | Human intervention, Bias, Discrimination, Trustworthy AI, Privacy, Data governance, Cybersecurity, Safety and security, AI standardisation |
| 04 | HLEG Sectoral considerations on policy and investment recommendations [16] | 8 | 12 | Health Care, Innovation, Risk, Copyright, Sector-specific AI, Trustworthy AI, Investment |
| 05 | EU AI Act [1] | 70 | 47 | Regulatory enforcement, Certification, Data protection, Monitoring, SME, Biometrics, Market surveillance, Traceability, Sandbox, AI value chain , General-purpose AI model, Harmonisation of rules, Discrimination, Documentation, Compliance, Risk, Cybersecurity, Oversight, Data governance, AI standardisation, Sector-specific AI |
| 06 | AI Act's prohibited AI guideline [3] | 62 | 33 | Sensitive data, Law enforcement, Surveillance, Privacy, Manipulation/deception/exploitation, Profiling, Criminal justice, Data protection, Emotion recognition, Biometrics, Transparency, Regulatory enforcement, Rule of law, Human autonomy, Vulnerable groups, National security, Psychological harm, Workers' rights, Social scoring |
| 07 | AI Act's general-purpose AI guideline [17] | 7 | 7 | Regulatory enforcement, AI value chain, Computational capabilities, Model training |
| 08 | AI Act's general-purpose AI code of practice [18] | 12 | 12 | Risk, Risk assessment, Safety and security, Risk mitigation, AI value chain , Documentation, Compliance, Copyright |

## Acknowledgements

This work has received funding from the European Commission's Horizon Europe Research and Innovation Programme under grant agreement No. 101177579 (FORSEE) and from Research Ireland at ADAPT, the Research Ireland Centre for AI-Driven Digital Content Technology at TCD and UCD #13/RC/2106_P2. For the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. We would like to thank Prof. John D. Kelleher and Dr. Vasudevan Nedumpozhimana for providing valuable advice on topic modelling.

## References


[1] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act); 2024. Available from: http://data.europa.eu/eli/reg/2024/1689/oj.

[2] European Commission and Directorate-General for Communications Networks, Content and Technology. Ethics Guidelines for Trustworthy AI. Publications Office of the European Union; 2019.

[3] European Commission. Commission Guidelines on Prohibited Artificial Intelligence Practices Established by Regulation (EU) 2024/1689 (AI Act); 2025.

[4] Clarke V, Braun V. In: Teo T, editor. Thematic Analysis. New York, NY: Springer New York; 2014. p. 1947-52. Available from: https://doi.org/10.1007/978-1-4614-5583-7_311.

[5] Papadopoulos T, Charalabidis Y. What do governments plan in the field of artificial intelligence? Analysing national AI strategies using NLP. Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance. 2020:100–111. Available from: https://doi.org/10.1145/3428502.3428514.

[6] Yang Q. LDA-based topic mining research on China's government data governance policy. Social Security and Administration Management. 2022;3(2).

[7] Wang S, Zhang Y, Xiao Y, Liang Z. Artificial intelligence policy frameworks in China, the European Union and the United States: An analysis based on structure topic model. Technological Forecasting and Social Change. 2025;212:123971.

[8] Roche C, Wall P, Lewis D. Ethics and Diversity in Artificial Intelligence Policies, Strategies and Initiatives. AI and Ethics. 2023;3(4):1095-115.

[9] Schiff D, Borenstein J, Biddle J, Laas K. AI Ethics in the Public, Private, and NGO Sectors: A Review of a Global Document Collection. IEEE Transactions on Technology and Society. 2021;2(1):31-42.

[10] Saheb T, Saheb T. Topical Review of Artificial Intelligence National Policies: A Mixed Method Analysis. Technology in Society. 2023;74:102316. Available from: https://www.sciencedirect.com/science/article/pii/S0160791X23001215.

[11] Saheb T, Saheb T. Mapping ethical artificial intelligence policy landscape: a mixed method analysis. Science and engineering ethics. 2024;30(2):9.

[12] Kajava K, Öhman E, Takagi NM, Nakajima-Wickham E, Vitiugin F. Post-GPT Policy: Risk and Regulation in EU AI Discourse. Proceedings of the International AAAI Conference on Web and Social Media. 2025;19(1):994-1006. Available from: https://ojs.aaai.org/index.php/ICWSM/article/view/35856.

[13] Suter V, Ma C, Pöhlmann G, Meckel M. When Politicians Talk AI: Issue-Frames in Parliamentary Debates Before and After ChatGPT. Policy & Internet. 2025;17(3).

[14] Commission E, Directorate-General for Communications Networks C, Technology, on Artificial Intelligence HLEG. Policy and investment recommendations for trustworthy AI. Publications Office of the European Union; 2019.

[15] European Commission and Directorate-General for Communications Networks, Content and Technology. The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment. Publications Office of the European Union; 2020.

[16] Commission E, Directorate-General for Communications Networks C, Technology. Sectoral Considerations on the Policy and Investment Recommendations for Trustworthy Artificial Intelligence. Publications Office; 2020.

[17] European Commission. Commission – Guidelines on the scope of the obligations for general-purpose AI models established by Regulation (EU) 2024/1689 (AI Act); 2025.

[18] General-Purpose AI Code of Practice; 2025. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai.

[19] Squires V. Thematic analysis. In: Varieties of qualitative research methods: Selected contextual perspectives. Springer; 2023. p. 463-8.

[20] Dhakal K. NVivo. Journal of the Medical Library Association: JMLA. 2022;110(2):270.

[21] Churchill R, Singh L. The Evolution of Topic Modeling. ACM Comput Surv. 2022 Nov;54(10s). Available from: https://doi.org/10.1145/3507900.

[22] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of machine Learning research. 2003;3(Jan):993-1022.

[23] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure; 2022. Available from: https://arxiv.org/abs/2203.05794.
[24] UNESCO. Recommendation on the Ethics of Artificial Intelligence; 2022.
[25] Pham C, Hoyle A, Sun S, Resnik P, Iyyer M. TopicGPT: A Prompt-based Topic Modeling Framework. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2024:2956-84.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251622*

# Foundations for Territorial Disambiguation in Law: A Preliminary Study Using the Education Act 2005

Safia KANWAL [a,1], Livio ROBALDO [b] and Hafsa DAR [b] and Davide LIGA [c] and Joseph ANIM [a]

[a] *Swansea University, UK*
[b] *University of Gujrat, PK*
[c] *University of Luxembourg, Luxembourg*

**Abstract.** In the devolved legal system of the United Kingdom (UK), legislative provisions may apply differently across regions such as England, Wales, Scotland, and Northern Ireland. Accurately determining this territorial scope is essential for legal interpretation and AI-assisted legal tools. However, metadata capturing jurisdictional applicability is inconsistently format, as only a few Acts include Territorial Application Annexes. This study presents a case study using the Education Act 2005 to evaluate the accuracy of automated methods for identifying territorial scope. We found that only 46.9% of sections matched in jurisdictional coverage. The best-performing approach achieved nearly 80% accuracy, showing that LLMs can effectively support scalable and explainable territorial disambiguation.

**Keywords.** Territorial Disambiguation, Legal—DocML, Large Language Models

## 1. Introduction

Legislation in the UK operates under a devolved structure, dividing authority between England, Wales, Scotland, and Northern Ireland. Determining where a provision applies—its territorial scope—is vital for legal interpretation and compliance automation. The National Archives provides statutes in *Akoma Ntoso XML (LegalDocML)* format, using ¡extent¿ tags to indicate territorial coverage. However, practical application may be narrower: a section marked as extending to England and Wales (E+W) may, in effect, apply only to England, depending on institutional references or legal context. This leads to ambiguity when `<extent>` alone is used for territorial inference. The *Education Act 2005*, which includes a detailed Territorial Application Annex, provides a unique ground truth to evaluate and improve automated territorial scope extraction using AI-based methods. As highlighted in the paper, many Acts lack such annexes, making territorial classification a manual and non-scalable task. This motivated the exploration of

[1]Corresponding Author: Livio Robaldo, livio.robaldo@swansea.ac.uk.

AI-driven inference techniques that can mirror human-style reasoning by analysing legal structure, content cues, and institutional references. [1]

## 2. Methodology

This study used the Education Act 2005 as a benchmark to evaluate automated approaches for territorial classification. Each section of the Act was extracted from its LegalDocML representation and paired with its corresponding ground-truth jurisdiction(s) from the Territorial Application Annex. The dataset was structured to include section text, hierarchical context (Part/Chapter titles), and ¡extent¿ metadata. Four progressively enriched methods were evaluated: (1) Extent-tag baseline, which directly mapped ¡extent¿ territories to predicted scope; (2) Named Entity Recognition (NER), which used jurisdictional mentions to refine classification; (3) LLM-based inference, where large language models (DeepSeek-R1 8B and 32B) were prompted using structured context to infer territorial scope; and (4) LLM with legal knowledge cues, which augmented prompts with domain-specific rules (e.g., references to S̈ecretary of Stateïmply England, Ëstynsuggests Wales). Model outputs were compared against the Annex, and accuracy was computed across all 128 sections to assess performance improvements across methods. All code and evaluation tools are publicly available[2].

## 3. Results and Discussion

The baseline extent tag approach achieved 46.9% accuracy, revealing that current metadata does not reliably represent territorial scope. NER improved performance to 61.7%, while the LLM-based methods achieved 63–73% accuracy. When enriched with legal knowledge, accuracy reached nearly 80%. These results demonstrate that combining structured context and simple legal cues can significantly improve AI-based territorial inference. However, manually embedding knowledge in prompts is not scalable, emphasizing the need for integrating legal knowledge graphs [2] in future research.

## 4. Conclusion and Future Work

Manual identification of territorial applicability in UK legislation is slow and challenging due to nuanced legal language and cross-references. This study shows that AI models can analyse statutory text to suggest likely territorial scope, reducing dependence on manually crafted annexes. Future work will focus on exploring more techniques to add external knowledge [3] and enhancing interpretability so that each AI-generated decision is supported with clear reasoning and can be easily verified.

## References

[1] Liga D, Robaldo F. Fine-tuning GPT-3 for legal rule classification using LegalDocML and Legal-RuleML. *Artif Intell Law*. 2023.

[2] https://github.com/SafiaK/legislation-territory-analyzer

[2] Cui J, Ning M, Li Z, Chen B, Yan Y, Li H, Ling B, Tian Y, Yuan L. Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model. 2024.
[3] Zhang Q, Dong J, Chen H, Zha D, Yu Z, Huang X. KnowGPT: Knowledge graph-based prompting for large language models. In: *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2024)*; 2024.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*

*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251623*

# Extracting Unwritten Sentencing Rules from Czech Higher Courts' Verdicts

Tomáš KNAP [a,1]
[a] *Faculty of Law, Charles University, Prague, Czech Republic*

**Abstract.** Czech courts operate without formal sentencing guidelines, relying on implicit principles that remain unmapped and inconsistently applied. This poster presents an early-stage project developing a four-phase pipeline to extract unwritten sentencing rules from Czech higher court decisions. The methodology combines: (i) thematic analysis for reasoning pattern identification, (ii) LLM-Based rule extraction and classification of rules (iii) cross-decision semantic matching and citation network analysis, and (iv) planned expert validation with criminal law scholars. The pipeline aims to reveal how principles emerge and propagate across the court hierarchy, from district courts to the Supreme Court. This work will provide the first systematic mapping of Czech implicit sentencing rules and a transferable methodology for other criminal law jurisdictions.

**Keywords.** Sentencing principles, higher courts, judicial discretion, legal argument mining, semantic role labeling, citation network analysis

## 1. Introduction and Motivation

Unlike continental jurisdictions with formal sentencing guidelines (e.g., Finland [1]), Czech higher courts lack explicit frameworks for principle-setting. While appeals are filed in approximately 10% of criminal cases, with courts modifying 25% of appealed decisions (primarily sentences), the Supreme Court generally does not review sentencing, leaving appellate courts as the de facto source of sentencing jurisprudence [2]. Yet research on post-communist Europe has focused on statutory frameworks rather than on how these courts actually create and apply rules, making systematic analysis essential to understand actual sentencing [3,2]. This project uses LLM-enhanced analysis to extract implicit sentencing principles from the decisions of Czech appellate courts, high courts, the Supreme Court, and the Constitutional Court, providing the first computational mapping of how principles are articulated and propagated across the court hierarchy.

## 2. Methods

As visualized in Figure 1, research is based on a four-phase interpretable pipeline that processes criminal sentencing decisions from Czech higher courts (Appellate Courts, High Courts, Supreme Court, and Constitutional Court) to identify and validate unwritten rules while maintaining transparency for legal experts:

[1]Corresponding author: tomas.knap@prf.cuni.cz

***Phase I: Thematic Analysis.*** The decision corpus is thematically analyzed to identify recurring reasoning patterns and normative themes. Structured prompting helps distinguish general principles from case-specific facts and outline initial candidate rules [4].

***Phase II: LLM-Based Rule Extraction and Classification.*** Instruction-tuned large language models transform identified themes into candidate IF–THEN rules. Through iterative prompt engineering, few-shot examples from domain-specific terminology (aggravating/mitigating circumstances, proportionality principle) enable the models to classify outputs as genuine principles or case-specific holdings [5].

***Phase III: Cross-Decision Analysis.*** Validated rules are traced across decisions via semantic similarity (sentence-BERT for Czech legal texts), citation-network analysis capturing "following" and "distinguishing" relations, and temporal tracking of doctrinal evolution [6,7].

***Phase IV: Expert Evaluation and Refinement.*** Criminal-law experts assess doctrinal accuracy, applicability, and consistency with the Criminal Code. Their feedback informs the iterative refinement and final synthesis of findings.

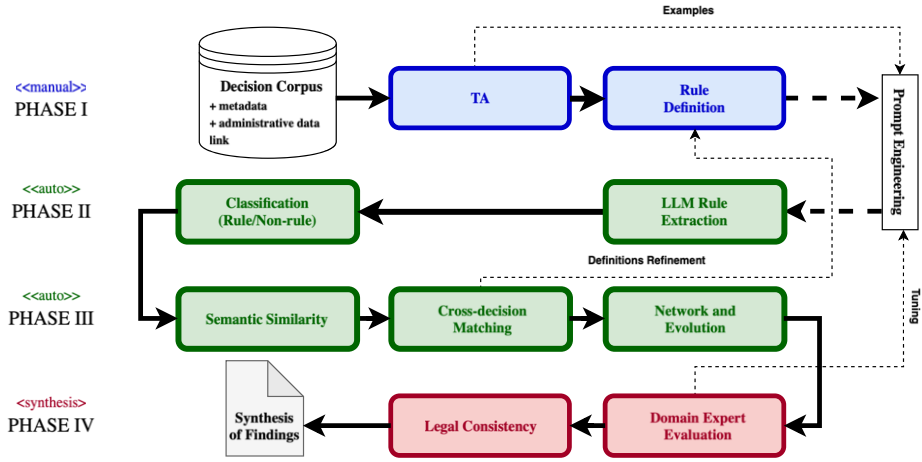



**Figure 1.** Methodology Architecture

## 3. Current Stage and Next Steps

This research is in its early stages, currently focused on corpus collection and annotation development. My immediate priorities include: (a) building a comprehensive corpus of Czech higher court sentencing decisions, (b) developing annotation guidelines for thematic analysis to identify reasoning patterns and normative themes, and (c) conducting pilot annotations to refine the taxonomy. Following this groundwork, I will test LLM-based rule extraction capabilities and assess the generalizability of extracted rules across different case types and courts using the annotated data. The pipeline described above represents my planned methodology, with phases I and II currently under active development.

## References

[1] Drápal J, Plesničar M. Sentencing Elsewhere: Structuring Sentencing Discretion in Post-communist Europe. European Journal on Criminal Policy and Research. 2023 Dec.
[2] Drápal J, Vanča T. In: Justifying Sentence Changes at Appellate Courts. 1st ed. London: Routledge; 2024. p. 63-79.
[3] Tata C. In: Sentencing Decision-Making: Unravelling the Enigma. Cham: Springer International Publishing; 2020. p. 1-11.
[4] Braun V, Clarke V. Using Thematic Analysis in Psychology. Qualitative Research in Psychology. 2006 Jan;3(2):77-101.
[5] Bakker R. Semantic Role Labelling for Dutch Law Texts.
[6] Novotna T, Harasta J, Kol J. Topic Modelling of the Czech Supreme Court Decisions. 2020.
[7] Sido J, Pražák O, Přibáň P, Pašek J, Seják M, Konopík M. Czert – Czech BERT-like Model for Language Representation. In: Proceedings of the Conference Recent Advances in Natural Language Processing - Deep Learning for Natural Language Processing Methods and Applications. INCOMA Ltd. Shoumen, BULGARIA; 2021. p. 1326-38.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251624*

# LeAK-GraphRAG: A Legal Academic Knowledge GraphRAG Framework with Hybrid Retrieval for Mitigating Hallucinations

Jia LIU[a,b,1] and Jianhao TANG[a,b]
[a] *Law School, Huazhong University of Science and Technology, Wuhan, Hubei, China*
[b] *Hubei Judicial Big Data Research Center, Wuhan, Hubei, China*

**Abstract.** This study introduces LeAK-GraphRAG, a framework that leverages knowledge graphs tailored for legal papers to enhance the semantic understanding of large language models (LLMs) regarding specialized legal academic knowledge and reduce hallucinations. Using a corpus of legal papers, we automatically extracted entities and relations with LLMs to construct a knowledge graph and systematically evaluated the performance of three retrieval methods and five LLMs. Experimental results demonstrated that the hybrid retrieval strategy achieved the best overall performance across most evaluation metrics, effectively combining semantic similarity with the structural relations of knowledge graph entities.

**Keywords.** GraphRAG, hybrid retrieval, legal information retrieval, knowledge graph, hallucinations

## 1. Introduction

In specialized areas of legal research, the knowledge provided by large language models (LLMs) remains fragmented, lacking domain-specific expertise. Therefore, LLMs frequently hallucinate when addressing specialized legal academic questions, often producing inaccurate, irrelevant, or fabricated information [1]. Retrieval-augmented generation (RAG) combines LLMs with external knowledge bases, improving the accuracy and relevance of generated content. Its effectiveness has also been validated in the legal domain [2,3,4]. On this basis, graph retrieval-augmented generation (GraphRAG) [5,6] integrates the relational reasoning capability of knowledge graphs with the semantic understanding capability of vector retrieval, further mitigating the hallucination issue in LLMs [7]. Considering that constructing knowledge graphs from domain-specific legal papers could enhance the output capabilities of LLMs in academic legal contexts, this study proposes a GraphRAG framework—LeAK-GraphRAG[2].

[1] Corresponding Author: Jia Liu, jialiu@hust.edu.cn. This research is supported by National Key Research and Development Program of Ministry of Science and Technology "Research on Legislation System for Trusted Artificial Intelligence" (No. 2022ZD0120100) and Interdisciplinary Research Support Program of Huazhong University of Science and Technology "Similar Case Recommendation Based on Enhanced Knowledge Graph Attention Network" (No. 2023JCYJ019).

[2] We share our code on github: https://github.com/vidal-77/LeAK-GraphRAG

## 2. Methods and Results

In the knowledge graph construction stage, we used the China National Knowledge Infrastructure as our data source and applied the following search formula: topic = (“training data” OR “dataset” OR “artificial intelligence training data”) AND (“legal regulation” OR “regulatory framework” OR “governance” OR “law” OR “risk”). This search retrieved 29 relevant papers, which were divided into text chunks for entity relation extraction, as shown in Figure 1. The resulting knowledge graph comprised 1163 entities and 1113 relations. Additionally, 1091 legal question-answer pairs were generated based on the extracted text chunks, serving as test datasets to support evaluation. Evaluation metrics included diversity, comprehensiveness, empowerment, directness, semantic similarity, and accuracy. The semantic similarity was computed and evaluated using BERT, while we used DeepSeek to determine the accuracy of the answers and invited legal experts to manually compare and verify the accuracy rate of each model using different retrieval methods. In the knowledge retrieval stage, we adopted three retrieval methods. Vector retrieval achieves semantic matching retrieval by calculating the cosine similarity between query text and the textual descriptions of entities and relationships in the knowledge graph. Cypher retrieval integrates semantic matching with graph structural features, calculating a hybrid score using vector similarity and graph relationship weights. Hybrid retrieval merges results from both methods. In the answer generation stage, we employed selected LLMs to generate answers for legal questions in the test dataset, combined with relevant texts obtained in the previous stage.

Experimental results indicated that hybrid retrieval achieved the highest scores for comprehensiveness, diversity, empowerment, and directness across the LLMs because hybrid retrieval combined the flexibility of textual semantics with the structure of graph relationships, compensating for the inadequacy of vector retrieval in handling structured relational information and overcoming the limitations of Cypher retrieval in capturing unstructured information. The semantic similarity scores and accuracy rates among the five LLMs were generally high, indicating that the answers generated by all models were closely aligned with the corresponding reference answers.

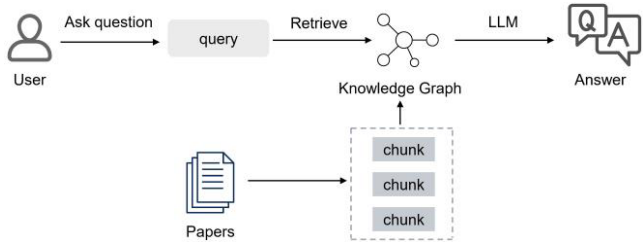



**Figure 1.** LeAK-GraphRAG Workflow Diagram

## 3. Conclusion

This study proposed LeAK-GraphRAG, a GraphRAG framework on legal academic knowledge. Evaluation results demonstrated that LeAK-GraphRAG could enhance the output quality of LLMs in the legal academic domain and reduce hallucinations.

## References

[1] Ji Z, Yu T, Xu Y, Lee N, Ishii E, Fung P. Towards mitigating LLM hallucination via self-reflection. In: Findings of the Association for Computational Linguistics: EMNLP 2023; 2023. p. 1827-43.

[2] Nguyen HT, Satoh K. ConsRAG: Minimize LLM Hallucinations in the Legal Domain. In: Jaromir Savelka, Jakub Harasta, Tereza Novotna and Jakub Misek, editors. Legal Knowledge and Information Systems - JURIX 2024: The Thirty-seventh Annual Conference, Brno, Czechia, 11-13 December 2024. vol. 95 of Frontiers in Artificial Intelligence and Applications. IOS Press; 2024. p.327-32. Available from: https://doi.org/ 10.3233/FAIA395.

[3] Ferraris AF, Audrito D, Siragusa G, Piovano A. Legal Chunking: Evaluating Methods for Effective Legal Text Retrieval. In: Jaromir Savelka, Jakub Harasta, Tereza Novotna and Jakub Misek, editors. Legal Knowledge and Information Systems - JURIX 2024: The Thirty-seventh Annual Conference, Brno, Czechia, 11-13 December 2024. vol. 95 of Frontiers in Artificial Intelligence and Applications. IOS Press; 2024. p.275-81. Available from: https://doi.org/ 10.3233/FAIA395.

[4] Do TPP, Cao NDD, Tran KQ, Nguyen KV, et al. R2GQA: retriever-reader-generator question answering system to support students understanding legal regulations in higher education. Artificial Intelligence and Law.2025 May:1-46, doi: https://doi.org/10.1007/s10506-025-09457-7.

[5] Edge D, Trinh H, Cheng N, Bradley J, Chao A, Mody A, et al. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. arXiv:240416130.2025.

[6] Sun J, Xu C, Tang L, Wang S, Lin C, Gong Y, et al. Think-on-Graph: Deep and Responsible Reasoning of Large Language Model on Knowledge Graph. arXiv:230707697.2024.

[7] Hu Y, Lei Z, Zhang Z, Pan B, Ling C, Zhao L. GRAG: Graph Retrieval-Augmented Generation. arXiv:240516506. 2024.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*

*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251625*

# A GraphRAG Approach to Enhancing the Ability of Causal Reasoning in Autonomous Driving Cases

Jia LIU[a,b,1] and Xingtong CHEN[a]
[a] *Law School, Huazhong University of Science and Technology, Wuhan, Hubei, China*
[b] *Hubei Judicial Big Data Research Center, Wuhan, Hubei, China*

**Abstract.** Large language models (LLMs) are significantly limited in causal reasoning in the legal domain. Determining legal liability in autonomous-driving traffic accidents inherently involves complex causal relationships, thereby providing a typical scenario for exploring and enhancing the causal reasoning capabilities of LLMs in legal contexts. In this study, GraphRAG is employed to provide more accurate and interpretable causal reasoning pathways for LLMs to determine legal liability, thereby enhancing their causal reasoning capabilities in the context of determining autonomous-driving traffic accident liability. We construct a knowledge graph from expert-annotated judicial documents using Neo4j and integrate it with LLMs using the LangChain framework. After referencing the legal causal reasoning paths retrieved from similar cases, LLMs could generate causal reasoning processes and final legal liability determinations for unresolved cases. This study could provide insights into judicial practices in autonomous-driving traffic accidents and advance the application of LLM-based causal reasoning in the legal domain.

**Keywords.** GraphRAG, causal reasoning, knowledge graph, autonomous driving, legal Q&A

## 1. Introduction

Large language models (LLMs) exhibit high hallucinations and poor interpretability in legal question-answering tasks [1]. These shortcomings are particularly pronounced in tasks that require legal causal reasoning [2]. Determining legal liability in autonomous-driving traffic accidents involves complex causal relationships [3], providing a typical scenario for investigating the causal reasoning capabilities of LLMs in the legal domain. Accordingly, this study aims to utilize GraphRAG to provide LLMs with more accurate and interpretable causal reasoning pathways for answering legal liability determination questions, thereby enhancing their causal reasoning capabilities in determining autonomous-driving traffic accident liability[4,5,6].

[1] Corresponding Author: Jia Liu, jialiu@hust.edu.cn. This research is supported by National Key Research and Development Program of Ministry of Science and Technology "Research on Legislation System for Trusted Artificial Intelligence" (No. 2022ZD0120100) and Interdisciplinary Research Support Program of Huazhong University of Science and Technology "Similar Case Recommendation Based on Enhanced Knowledge Graph Attention Network" (No. 2023JCYJ019).

## 2. Methods

We use publicly available judicial documents on autonomous-driving traffic accidents from China Judgments Online as the foundational corpus. These documents contain comprehensive information, including case facts, adjudication bases, adjudication reasoning, and adjudication results. Using techniques such as knowledge extraction and fusion, we construct a knowledge graph using Neo4j. To obtain more accurate and comprehensive information that better aligns with legal logic [7,8], we integrate relevant laws and regulations [9] applicable to autonomous-driving cases and rely on legal experts to annotate legal causal reasoning paths for cases in the knowledge graph. Annotations include legal causal types (e.g., direct and indirect causation) and legal causal relationship patterns (e.g., single cause-single effect and multiple cause-multiple effects). We construct a test dataset containing information on autonomous-driving cases and questions regarding legal liability determination, organizing legal experts to answer these questions. The cases in the test set cover complex scenarios involving multiple parties, behaviors, and causal relationships potentially influenced by factors such as vehicle malfunctions and human intervention errors. Based on LangChain, LLMs are integrated with the knowledge graph. LLMs retrieve relevant information (e.g., legal causal reasoning paths) from cases similar to the unresolved case in the knowledge graph [10]. By referencing legal causal reasoning paths from similar cases, LLMs generate these paths for unresolved cases and answer questions related to legal liability determination [11]. The overall process is illustrated in Figure 1. To further enhance LLM interpretability, our improved approach offers users traceable similar case-matching analysis processes, causal reasoning paths that adhere to legal logic, and reliable liability determination results when answering questions. The results obtained by the proposed approach will be evaluated against those generated by only LLMs, both quantitatively and qualitatively, using the answers provided by legal experts as reference answers. Specifically, the causal reasoning capabilities of the proposed approach will be assessed from aspects such as comprehensiveness, interpretability, and accuracy.

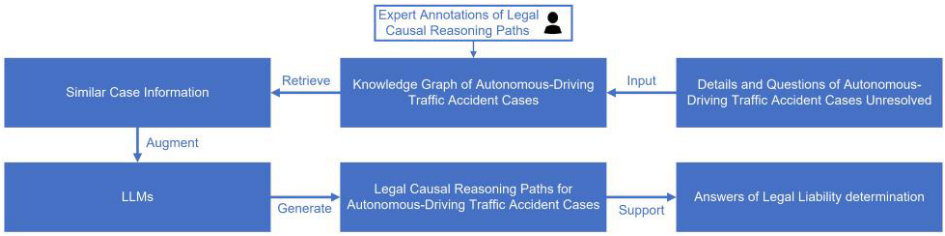



**Figure 1.** Legal liability determination process in autonomous-driving cases

## 3. Conclusion

In this study, GraphRAG is applied to enhance the accuracy and interpretability of LLMs for legal causal reasoning tasks in a specific legal scenario. This approach could offer insights into judicial practices in autonomous-driving traffic accidents and advance the application of LLM-based causal reasoning in the legal domain.

## References


[1] Dimos J, Fourel E. Legal Question Answering Based on Logical Reasoning in Large Pretrained Language Models. In: J. Savelka et al., editors. Proceedings of the 37th International Conference on Legal Knowledge and Information Systems (JURIX); 2024 Dec 11-13; Brno, Czech Republic: IOS Press; c2024. p. 366-368, doi: 10.3233/FAIA241271.

[2] Kang X, Qu L, Soon L, Li Z, Trakic A. Automating IRAC Analysis in Malaysian Contract Law using a Semi-Structured Knowledge Base. Artificial Intelligence and Law. 2025 Jul; doi: 10.1007/s10506-025-09467-5.

[3] Liepiņa R, Wyner A, Sartor G, Lagioia F. Argumentation Schemes for Legal Presumption of Causality. Proceedings of the 19th International Conference on Artificial Intelligence and Law (ICAIL); 2023 Jun 19 - 23; Braga, Portugal: ACM Press; c2023. p. 157-166, doi: 10.1145/3594536.3595130.

[4] Cai T, Liu Y, Zhou Z, Ma H, Zhao Z, Wu Z, et al. Driving with Regulation: Interpretable Decision-Making for Autonomous Vehicles with Retrieval-Augmented Reasoning via LLM. arXiv preprint arXiv: 241004759.2024.

[5] Samarajeewa C, De Silva D, Osipov E, Alahakoon D, Manic M. Causal Reasoning in Large Language Models using Causal Graph Retrieval Augmented Generation. Proceedings of the 16th International Conference on Human System Interaction (HSI); 2024 Jul 8-11; Paris, France: IEEE Access; c2024. p. 1-6, doi: 10.1109/HSI61632.2024.10613566.

[6] Wang N, Han X, Singh J, Ma J, Chaudhary V. CausalRAG: Integrating Causal Graphs into Retrieval-Augmented Generation. In: Che W, Nabende J, Shutova E, Pilehvar MT, editors. Findings of the Association for Computational Linguistics: ACL; 2025 Jul 2-Aug 1; Vienna, Austria: Association for Computational Linguistics; c2025. p. 22680–22693, doi: 10.18653/v1/2025.findings-acl.1165.

[7] Xiong M. On the Reference Rules in Legal Logic. Social Sciences in China. 2009 Aug; 30(3): 58-74, doi: CNKI: SUN: SHEK.0.2009-03-009.

[8] Liepiņa R, Sartor G, Wyner A. Arguing About Causes in Law: A Semi-formal Framework for Causal Arguments. Artificial Intelligence and Law. 2020 Mar; 28(1): 69-89, doi:10.1007/s10506-019-09246-z.

[9] Standing Committee of the National People's Congress. Road Traffic Safety Law of the People's Republ -ic of China; 2021. Accessed: 23rd August 2025. Available from: http://www.npc.gov.cn/npc///c2/c12435/c12488/201905/t20190522_70904.html.

[10] Ren T, Zhang Z, Jia B, Zhang S. Retrieval-Augmented Generation-aided causal identification of aviation accidents: A large language model methodology. Expert Systems with Applications. 2025 Jun; 278:127306, doi: 10.1016/j.eswa.2025.127306.

[11] Xiong M. The Modelling Reasoning Approaches in Artificial Intelligence and Law. Seeking Truth. 2020 Jun; 47(6):89-100, doi: 10.19667/j.cnki.cn23-1070/c.2020.06.010.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251626*

# Automated Identification of Criminal Offences from Case Descriptions Using Ontologies, Rules and LLMs

Jelena MATKOVIĆ [a], Marko MARKOVIĆ [a] and Stevan GOSTOJIĆ [a]
[a] *University of Novi Sad, Faculty of Technical Sciences, Serbia*

**Abstract.** This paper presents a system that automatically identifies criminal offences from Montenegrin court decisions. Factual attributes are extracted using LLMs and mapped to an OWL ontology with SWRL rules to infer applicable legal provisions and sanctions. Evaluation conducted on a dataset of court decisions demonstrates consistently high levels of accuracy and recall, confirming the feasibility of integrating LLM-based factual extraction with ontology-driven legal reasoning.

**Keywords.** rule-based reasoning, prompt engineering, automated legal reasoning, criminal law, legal ontology

## 1. Introduction

Automatically identifying criminal offences from unstructured legal narratives is challenging due to complex language and variability in narrative expression of factual descriptions. Previous work has demonstrated that ontology-based reasoning can support automated decision-making in the criminal domain [1,2]. In this study, we integrate large language model extraction with an OWL ontology [3] and SWRL rules [4] to identify theft-related offences (Articles 239–242 of the Criminal Code of Montenegro [5]) from Montenegrin court decisions. ChatGPT [6] is used to extract factual attributes, which are instantiated within the ontology.

## 2. Methodology

Our dataset consists of 2,357 basic court decisions from the official judiciary portal [7], of which 841 contained at least one theft-related offence. The system represents legal knowledge using a three-layer ontology comprising the Legal Norm Layer, which encodes articles and their hierarchical structure; the Legal Case Layer, which captures entities such as offenders, victims, and offence elements (e.g., intent, execution, and circumstances); and the SWRL Layer, containing 42 rules that infer applicable legal provisions based on instantiated facts. The ontology and rules are available online [8]. Factual attributes are extracted from unstructured case descriptions using 28 prompt questions posed to ChatGPT. Each answer is mapped to a triplet in the ontology; for exam-

ple, a response of "Yes" to "Did the accused use force?" is represented as (`Offender`, `hasExecution`, `Force`). Rule-based reasoning then applies SWRL rules to determine the relevant legal articles. A sample rule for theft (Article 239, Paragraph 1) can be expressed as:

```
Case(?case) ^ hasOffender(?case, ?offender) ^
hasIntent(?offender, UnlawfulGain) ^ hasStolen(?offender,
?goods) ^ StolenGood(?goods) ->
ruleBasedInferred(?case, Article_239_Para_1)
```

## 3. Results

We distinguish six categories to match the system output and the qualifications found in the decisions, with the overall performance summarised in Table 1.

**Table 1.** Evaluation results by category

| Category | Count | Percentage (%) | Mapped as |
|---|---|---|---|
| Absolute match (article+paragraph+point) | 493 | 20.92 | TP |
| Match + additional articles suggested | 286 | 12.13 | TP |
| No articles expected or detected | 1333 | 56.55 | TN |
| No articles expected, some detected | 183 | 7.76 | FP |
| Articles expected, detected none | 53 | 2.25 | FN |
| Partially recognised | 9 | 0.38 | FN |
| Accuracy = 89.6%, Precision = 81.0%, Recall = 92.6%, F1-score = 86.4% | | | |

The classification metrics by offence type are shown in Table 2.

**Table 2.** Classification metrics by offence type

| Crime | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Theft | 0.899 | 0.683 | 0.944 | 0.791 |
| Aggravated Theft | 0.938 | 0.729 | 0.965 | 0.832 |
| Robbery Theft | 0.996 | 0.308 | 0.800 | 0.444 |
| Robbery | 0.938 | 0.133 | 0.957 | 0.233 |

## 4. Discussion and Conclusion

Combining LLM-based extraction with ontology and SWRL reasoning achieves high recall and reliable classification of theft-related offences. Most errors arise from semantic ambiguity and overgeneralised LLM outputs. The system demonstrates feasibility as a decision-support tool for preliminary legal qualification. Future work will expand the ontology to additional offences and explore local LLM deployment to preserve sensitive data.

**Acknowledgements.** This research has been supported by the Ministry of Science, Technological Development and Innovation (Contract No. 451-03-137/2025-03/200156) and the Faculty of Technical Sciences, University of Novi Sad through project "Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad 2025" (No. 01-50/295).

## References

[1] El Ghosh M, Naja H, Abdulrab H, Khalil M. Towards a legal rule-based system grounded on the integration of criminal domain ontology and rules. Procedia Computer Science. 2017;112:632-642. doi:10.1016/j.procs.2017.08.109.

[2] Navarrete F, Garrido AL, Bobed C, Atencia M, Vallecillo A. Ontology-driven automated reasoning about property crimes. Business & Information Systems Engineering. 2024;1-24. doi:10.1007/s12599-024-00886-3.

[3] OWL 2 Web Ontology Language Document Overview (Second Edition). W3C Recommendation, 2012. Available at: https://www.w3.org/TR/owl2-overview/.

[4] SWRL: A Semantic Web Rule Language Combining OWL and RuleML. W3C Member Submission, 2004. Available at: https://www.w3.org/Submission/SWRL/.

[5] Criminal Code of Montenegro. Official Gazette of Montenegro. Available at: https://www.sluzbenilist.me/propisi/E6AAA874-5381-42ED-A8A9-35C00E1998C6.

[6] OpenAI. ChatGPT API. Available at: https://platform.openai.com.

[7] The Courts of Montenegro. Decisions Database. 2025. Available at: https://sudovi.me/sdvi/odluke.

[8] Matkovic J, Markovic M, Gostojic S. Ontology for Identification of Criminal Offences. Available at: https://github.com/legal-informatics/ontology-for-identification-of-criminal-offences.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*

*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251627*

# Principles of Law: LLMs vs RegEx

Marianna MOLINARI [a,b,c,1], Ilaria Angela AMANTEA [b], Marinella QUARANTA [a,b,c] and Guido GOVERNATORI [d]

[a] *University of Bologna - LaST-JD, Legal Studies Dept.*
[b] *University of Turin - Computer Science Dept.*
[c] *Vrije Universiteit Brussel - PREC Dept.*
[d] *Central Queensland University*

**Abstract.** Principles of law (PoLs) offer a framework for case interpretation and ensure legal consistency, especially in areas lacking specific legislation. Courts use PoLs to establish binding precedents that guide decision-making. However, PoLs are expressed in a non-standard way in rulings, making them difficult to identify at a glance without interpretative reading. Our study tested ChatGPT, Claude, and DeepSeek in extracting PoLs, verifying their results with Regular Expressions (RegEx). While Large Language Models (LLMs) showed promise, their outputs required significant expert oversight, revealing limitations in legal research, particularly in terms of accuracy and quantity.

**Keywords.** Principles of Law, LLMs, RegEx, Extraction.

## 1. Introduction and Methodology

When judges encounter new demands not addressed by existing laws, they create PoLs through their decisions [1,2]. Once stated or reaffirmed in a judgment, PoLs become precedents [3]. Precedents promote legal certainty and fairness, particularly when laws are absent or unclear [4,5]. Analyzing precedents to identify the core PoL(s) is crucial, as judgments are typically written in various styles that affect how PoLs and related citations are presented, making prompt identification challenging [6]. After addressing the issue [7,8], we focused on automatically extracting PoLs using ChatGPT [9], comparing its performance with RegEx [10,11]. This paper extends the comparative analysis by including various LLMs and conducts a retrospective evaluation with a RegEx baseline, generated by ChatGPT and applied via a simple text editor. This approach guarantees that the process can be reproduced and verified for falsifiability [12,13]. As tested, we prompted ChatGPT to generate a Python script [14] using RegEx to extract paragraphs with quoted text and specific keywords [15].The script processed .docx files, extracted PoLs, and outputted a separate CSV file for each document. PoLs were extracted from 29 randomly selected Italian judgments in .docx format from the De Jure database. The judgments were divided into two sets: originals, annotated manually by three legal experts, and copies, processed using LLMs and RegEx. Annotation followed guidelines identifying: explicit direct PoLs, clearly stated in quotation marks, often starting with an introductory phrase, and explicitly connected to a prior citation; explicit indirect PoLs,

[1]Corresponding Author. Email: marianna.molinari2@unibo.it.

paraphrased versions of direct PoLs, presented in various styles, and also explicitly linked to a prior citation; and implicit PoLs, inferred from context, not explicitly stated, and paraphrased in different ways[2].

## 2. Findings and Results

Table 1 presents the extraction results for ChatGPT, Claude and DeepSeek[3]:

| | Total PoLs | Implicit | Direct | Indirect | Hallucination | Not-PoL | PoL not found by annotators |
|---|---|---|---|---|---|---|---|
| **Annotators** | 395 | 49 | 173 | 173 | 0 | 0 | 0 |
| **DeepSeek** | 179 | 14 | 98 | 67 | 3 | 27 | 3 |
| **Claude** | 126 | 9 | 70 | 47 | 0 | 18 | 5 |
| **ChatGPT** | 74 | 5 | 42 | 27 | 4 | 19 | 2 |
| **RegEx** | 188 | 6 | 129 | 53 | 0 | 49 | 0 |

**Table 1.** Details of the experiments.

The corresponding confusion matrices have been calculated, as shown in Table 2:

| Annotators | | | DeepSeek | | | Claude | | | ChatGPT | | | RegEx | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Yes | No | | Yes | No | | Yes | No | | Yes | No | | Yes | No |
| PoLs | 395 | 7 | PoLs | 179 | 223 | PoLs | 126 | 276 | PoLs | 74 | 328 | PoLs | 188 | 214 |
| not PoLs | - | - | not PoLs | 30 | - | not PoLs | 18 | - | not PoLs | 23 | - | not PoLs | 49 | - |

**Table 2.** Confusion Matrices.

The precision, recall, accuracy, and F1-score are summarized as follows:

| | Annotators | DeepSeek | Claude | ChatGPT | RegEx |
|---|---|---|---|---|---|
| **Precision** | 0.982 | 0.445 | 0.313 | 0.184 | **0.468** |
| **Recall** | 1.0 | 0.856 | **0.875** | 0.723 | 0.793 |
| **Accuracy** | 0.982 | 0.414 | 0.300 | 0.174 | **0.417** |
| **F1-score** | 0.991 | 0.586 | 0.461 | 0.296 | **0.588** |

The number of PoLs identified and their completeness or partiality are as follows:

- Annotators (Total PoLs: 395): 395 fully, 0 partially;
- DeepSeek (Total PoLs: 179): 150 fully, 29 partially;
- Claude (Total PoLs: 126): 111 fully, 15 partially;
- ChatGPT (Total PoLs: 74): 43 fully, 31 partially;
- RegEx (Total PoLs: 188): 187 fully, 1 partially.

## 3. Conclusions

This study highlights that while LLMs can provide a general overview of judgments, they remain susceptible to bias, hallucinations, and inaccuracies. These issues are especially problematic when dealing with PoLs, which can be mixed, reported partially, or confused with passages that are technically correct but do not constitute actual PoLs (Not-PoLs). Since the validity of PoLs depends on accuracy, it is crucial for domain experts to remain involved in overseeing their extraction.

[2] All rights reserved by Giuffrè Francis Lefebvre: the publisher provided consent for the use of the data in the present research only. Therefore, the text of the judgments cannot be shared.

[3] ChatGPT (September 2024), Claude (3.7 Sonnet), and DeepSeek (April 2025).

## References


[1] De Amicis G. La formulazione del principio di diritto e i rapporti tra Sezioni semplici e Sezioni Unite penali della Corte di Cassazione. Dir Pen Contemp. 2019;4:1-29.

[2] Scoditti E. Giurisdizione per principi e certezza del diritto. Questione Giustizia. 2018.

[3] Spada I, Mignone R, Molinari M, Audrito D, Quaranta M, Amantea IA, et al. Adopting Video-Recordings to Extract Embedded Domain Knowledge and Automate Legal Case Reading. In: Proc Workshop on Annotation of Legal Data – JURIX (In Press); 2023. .

[4] Maltz E. The nature of precedent. NCL Rev. 1987;66:367.

[5] Alexy R. On the structure of legal principles. Ratio Jur. 2000;13(3):294-304.

[6] Molinari M, Quaranta M, Amantea IA. Principles of Law: the Missing Link Between Cases and Rules. In: Proc ALP – JURIX (In Press); 2023. .

[7] Molinari M, Bonfanti C, Amantea IA. Principles of law: Approaching a Functional Extraction. In: Proc AI4Legs2023 – ICAIL (In Press); 2023. .

[8] Molinari M, Quaranta M, Amantea IA. How do Case Law and Principles of Law interact, Computationally? In: Proc Workshop on Innovation and Digitization of the Justice System – itAIS; 2023. p. 111.

[9] Molinari M, Quaranta M, Amantea IA, Governatori G. Using ChatGPT to extract principles of law for the sake of prediction: an exploration conducted on Italian judgments concerning LGBT(QIA+) rights. In: Proc AI4A2J2024 – JURIX (In Press); 2024. .

[10] Zeugmann T, Poupart P, Kennedy J, Jin X, Han J, Saitta L, et al. Precision and recall. Encyclopedia Mach Learn. 2011:781-1.

[11] Van Rijsbergen CJ. Information retrieval. 2nd ed. London: Butterworths; 1979.

[12] Popper KR. In: Science as falsification. Routledge; 1963. p. 33-9.

[13] Aho AV. Algorithms for finding patterns in strings. Algorithms Complex. 2014;1:255.

[14] Pedregosa F. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825-30.

[15] Molinari M, Amantea IA, Quaranta M, Governatori G. ChatGPT Unveils Its Limits: Principles of Law Deliver Checkmate; 2025. http://arxiv.org/abs/2510.19261.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251628*

# Packaging Thematic Analysis as an AI Workflow for Legal Research

Vojtěch POUR [a,1,2]

[a] *Faculty of Law, Charles University, Prague, Czech Republic*

**Abstract.** This work-in-progress paper introduces a reusable system that operationalises thematic analysis for legal research as a reproducible, agentic workflow. Researchers upload a JSON corpus keyed by case identifiers; plain-language intent (scope, granularity, focus, presentation) is mapped to a fixed sequence of phases orchestrated from the frontend. A lightweight web UI streams intermediate outputs and offers optional checkpoints with reversible edits. The workflow is model-agnostic, multilingual, and stabilises codebooks and candidate themes without ad-hoc prompt retuning on representative case-law datasets. Concretely, I package and expose in a frontend the approach explored by Drápal, Westermann, and Šavelka and collaborators. The current prototype is fully operational through candidate theme generation (Phase 3); the final consolidation and assignment stages (Phases 3b–4) are planned for completion by JURIX. Overall, the system lowers time-to-first-theme, improves refinement safety, and increases portability across jurisdictions and languages.

**Keywords.** Thematic analysis, legal research, LLMs, qualitative coding, agentic workflows, case law, NLP

## 1. Introduction and Motivation

Thematic analysis (TA) identifies and interprets patterns of meaning in text [1]. In legal research it supports close reading of case law to surface codes, propose themes, and map arguments [2], complementing information extraction of court-established facts in criminal-court opinions [3]. Manual TA is slow and hard to port across corpora and languages [4]. I therefore target a practical metric—*time-to-first-theme (TTF-Theme)*—to free expert time for curation while retaining approval.

I package the approach explored by Drápal, Westermann, and Šavelka into a frontend-orchestrated, instruction-responsive workflow with explicit human-in-the-loop checkpoints and audited, rule-based edits [2,5]. The design aligns with prior strands on segmentation, fast similarity search, cross-lingual transfer, and zero-shot legal annotation [5,6,7]. Our model-agnostic, multilingual architecture reduces bespoke prompt engineering and increases reproducibility across jurisdictions.

[1] Corresponding author: vojtech.pour@gmail.com
[2] Code and demo: https://github.com/vojtechpour/typical_crimes_jurix.git

## 2. Workflow, UI, and Human-in-the-Loop

*Overview.* A frontend orchestrates TA phases P2→P3→P3b→P4 with lightweight checks, explicit checkpoints, and cross-document aggregation to stabilise codebooks and candidate themes. The pipeline operationalises the Drápal–Westermann–Šavelka approach in a UI with reversible edits and consistent exports.

*Modules.* Five modules correspond to TA phases:

1) **Data & Theme Explorer (pre-flight):** upload a single JSON keyed by `case_id`; preview schema and per-case text; a quality panel flags duplicates, missing fields, and token-length outliers.
2) **Initial Codes (Phase 2):** per-case code pills with *Regenerate*/*Add*/*Edit*; optional user-supplied seed codes; adjustable code granularity; quick CSV/HTML exports and progress tracking.
3) **Candidate Themes (Phase 3):** cluster-and-label codes across cases; themes stream progressively with short descriptions and evidence snippets; automatic consolidation with optional user regrouping.
4) **Final Themes (Phase 3b):** *planned* merge/split/rename/move operations with audit-tracked edits; finalize the theme set.
5) **Assign Final Themes (Phase 4):** *planned* application of the final theme set across all cases with thresholds and summaries; export enriched JSON and compact reports.

*User control over AI outputs.* Users can shape AI behaviour through a model selector (options including the GPT-5, Claude, and Gemini families), quick instruction pills and a free-form instruction box in Phase 2; per-case *Regenerate* and *Add more codes* dialogs accept custom instructions; Phase 3 also exposes the model selector; and an optional Theme Assistant panel supports conversational requests with provider-dependent toggles for reasoning effort and verbosity.

The system expects a single JSON object keyed by case identifiers. On upload, the explorer validates the schema and reports issues. Users may begin at any phase and may bring annotated or synthetic data for earlier steps (for example, jumping directly to P3 with an existing codebook). As of October 2025, the prototype is operational through P3; P3b–P4 and finer UI controls are scheduled for JURIX. Looking ahead, an agentic controller will run phases independently, iterate on subsets, edit specific codes or themes under constraints, and execute analyses over the JSON while reusing the same action schema and audit trail.

## 3. Conclusion and Artifact Availability

I package TA for legal research as an instruction-responsive, end-to-end workflow with checkpoints, audited edits, and traceable exports. The system lowers time-to-first-theme while preserving methodological control and portability across jurisdictions and languages. The repository with code, prompts, configuration, and sample datasets is available at https://github.com/vojtechpour/typical_crimes_jurix.git; a demo will be shown at JURIX.

## References

[1] Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol*. 2006;3(2):77–101. doi:10.1191/1478088706qp063oa.

[2] Drápal J, Westermann H, Šavelka J. Using large language models to support thematic analysis in empirical legal studies. *arXiv preprint* arXiv:2310.18729; 2023. doi:10.48550/arXiv.2310.18729.

[3] Bendová K, Knap T, Černý J, Pour V, Šavelka J, Kvapilíková I, et al. What Are the Facts? Automated Extraction of Court-Established Facts from Criminal-Court Opinions. In: *Proceedings of the ASAIL 2025 Workshop at ICAIL*. Northwestern Pritzker School of Law, Chicago (IL), USA; 2025.

[4] Braun V, Clarke V. Reflecting on reflexive thematic analysis. *Qual Res Sport Exerc Health*. 2019;11(4):589–597. doi:10.1080/2159676X.2019.1628806.

[5] Šavelka J, Ashley KD. Segmenting U.S. court decisions into functional and issue specific parts. In: Palmirani M, editor. *Legal Knowledge and Information Systems – JURIX 2018*. (Frontiers in Artificial Intelligence and Applications; vol. 313). Amsterdam: IOS Press; 2018. p. 111–120. doi:10.3233/978-1-61499-935-5-111.

[6] Westermann H, Šavelka J, Walker VR, Ashley KD, Benyekhlef K. Sentence embeddings and high-speed similarity search for fast computer-assisted annotation of legal documents. In: *Legal Knowledge and Information Systems – JURIX 2020*. (Frontiers in Artificial Intelligence and Applications; vol. 334). Amsterdam: IOS Press; 2020. p. 164–173. doi:10.3233/FAIA200860.

[7] Šavelka J, Ashley KD. The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts. *Front Artif Intell*. 2023;6:1279794. doi:10.3389/frai.2023.1279794.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*

*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251629*

# Plans and Diversions EA

Galileo SARTOR [a,1], Guido GOVERNATORI [b], Giuseppe PISANO [c], Antonino ROTOLO [c], and Adam WYNER [a]
[a] *Swansea University, United Kingdom*
[b] *Central Queensland University, Australia*
[c] *University of Bologna, Italy*

**Abstract.** Autonomous vehicles (AVs) must abide by the Highway Code. An AV agent would make plans which incorporate legal reasoning. Such plans must account for contingencies wherein the agent chooses between following a current plan that would lead to a violation of the law or providing an alternative plan which would not lead to a violation - a diversion. The paper utilises Defeasible Deontic Logic representing norms and encoded in Answer Set Programming (ASP) which is integrated with planning in ASP. The paper introduces diversions to address contingencies with respect to legal reasoning.

**Keywords.** defeasible deontic logic, answer set programming, planning, autonomous vehicles

## 1. Introduction

Planning is critical in many domains in order to attain goals. However, contingencies may arise, leading to changes in plans. We refer to such variant plans as *diversions*. Planning in a legal setting introduces legal norms – obligations, prohibitions, or permissions – on actions along with sanctions and reparations where the norms are violated. An agent's plan and behaviour may vary depending on its *disposition* towards legal risk or penalty. The research focuses on plans of *autonomous vehicles* (AVs) which have to abide by the Highway Code[2].

Taking a logic-based approach, the research builds on an encoding of Defeasible Deontic Logic (DDL), which represents and reasons with legal knowledge, in Answer Set Programming [1], and utilising ASP planners – *DDL-ASP planner*. The contribution enables the DDL-ASP planner to reason with contingencies and propose diversions. The research is scoped: one type of legal signal - the stoplight; one prohibition (on crossing a red light); dispositions of *agens legalis* and *agens economicus*; no reasoning behind the type of agent; limited cost/benefit analysis; a small environment; and no mitigations. The system is the basis for future work.

## 2. Defeasible Deontic Logic, ASP, and Planning

Defeasible Deontic Logic (DDL) is an efficient rule-based, non-monotonic formalism for legal reasoning [2]. It combines features of Defeasible Logic for the natural and efficient

[1]Corresponding Author: Galileo Sartor, galileo.sartor@swansea.ac.uk.
[2]The source code for this paper is available at https://github.com/Gilbocc/asp-av-planning

modelling of exceptions with deontic logic, including deontic operators (obligation and permission), and the ability to reason with violations and compensatory obligations. Answer Set Programming (ASP) has been established as a powerful formalism for declarative problem-solving [3]. DDL has been encoded in ASP [1]. And ASP has long incorporated planning [4]. The reader is referred to the literature for technical details.

## 3. Autonomous Vehicles

The complexity of autonomous driving motivates a hierarchical planning architecture which distinguishes between global, strategic route selection and local, context-aware decision making [5,6]. A global planner cannot account for contingencies, e.g., status of traffic lights, other road users, etc.; a local planner has no strategic reasoning [7]. Planning in AVs is generally limited to trajectory prediction [8,9]. There are different logic-based representations of traffic rules, e.g., as BDI models [10,11] or controlled natural languages [12].

In this context the *global* and *local* planner share the basic world model, differing in their access to the dynamic components, operating respectively under uncertainty (optimizing the plan by estimated time) and with access to the perceived state (optimizing also by norm compliance). The dynamic information is included in the ASP theory for the local planner before the grounding phase. This includes: the updated position, the current direction, the window size, the goal, and any mitigating circumstances such as being in an emergency.

We use the ASP optimization mechanisms so the local planner can optimize in different ways, depending on the explicit optimization criteria: minimising the number of violations; minimising the time to reach the goal; or some combination. In the Python code, the optimization is set when calling the local planner. While in the current state the reasoning with violations and optimisation is limited. In future developments, the reasoning will be made more context dependent.

The agent operates under a set of norms that regulate its behaviour[13]. A prescriptive rule obligates the agent to stop at red traffic lights, with a monetary fine in cases of non-compliance. A constitutive rule identifies the conditions under which we have a failure to stop, namely, when the agent is obligated to stop yet acts. Exceptions are modelled using a permissive rule that allows the agent to disregard red lights during emergency situations. In the short planner, where the execution of an action in a state were to lead to a violation, we say that the plan contains a violation. In such a case, a new long term plan may be calculated.

The environment is modelled as a directed graph consisting of nodes, some of which contain traffic lights. The agent can move between adjacent nodes, with actions modelled in terms of pre- and post-conditions, or stop in the current node. Where the action is taken, the agent's new position and direction are updated.

If the local plan is compliant (i.e., no violation would occur), it is executed, and the local plan is called on the next state. Otherwise, the global planner is invoked to find alternative paths to the global goal.

The analysis illustrates the benefits, in terms of simplicity and consistency, of a shared DDL-ASP logical representation which integrates the environment, actions, planning, and with reasoning about norms.

## References


[1] Governatori G. An ASP Implementation of Defeasible Deontic Logic. Künstliche Intell. 2024;38(1-2):79-88.

[2] Governatori G, Rotolo A. Defeasible Logic: Agency, Intention and Obligation. In: Lomuscio A, Nute D, editors. 7th International Workshop on Deontic Logic in Computer Science. No. 3065 in LNAI. Berlin: Springer; 2004. p. 114-28.

[3] Eiter T, Ianni G, Krennwallner T. In: Tessaris S, Franconi E, Eiter T, Gutierrez C, Handschuh S, Rousset MC, et al., editors. Answer Set Programming: A Primer. Berlin, Heidelberg: Springer Berlin Heidelberg; 2009. p. 40-110. Available from: https://doi.org/10.1007/978-3-642-03754-2_2.

[4] Son TC, Pontelli E, Balduccini M, Schaub T. Answer Set Planning: A Survey. Theory Pract Log Program. 2023;23(1):226-98.

[5] Maierhofer S, Moosbrugger P, Althoff M. Formalization of Intersection Traffic Rules in Temporal Logic. In: 2022 IEEE Intelligent Vehicles Symposium (IV); 2022. p. 1135-44. Available from: https://ieeexplore.ieee.org/document/9827153.

[6] Rakow A, Schwammberger M. Brake or Drive: On the Relation Between Morality and Traffic Rules when Driving Autonomously. In: Software Engineering 2023 Workshops; 2023. p. 104. Available from: https://publikationen.bibliothek.kit.edu/1000169851.

[7] Prakken H. On the problem of making autonomous vehicles conform to traffic law. Artif Intell Law. 2017;25(3):341-63.

[8] Manas K, Paschke A. Knowledge Integration Strategies in Autonomous Vehicle Prediction and Planning: A Comprehensive Survey. CoRR. 2025;abs/2502.10477. Available from: https://doi.org/10.48550/arXiv.2502.10477.

[9] Lin J, Zhou W, Wang H, Cao Z, Yu W, Zhao C, et al. Road Traffic Law Adaptive Decision-making for Self-Driving Vehicles. In: 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC); 2022. p. 2034-41. Available from: https://ieeexplore.ieee.org/document/9922208.

[10] Collenette J, Dennis LA, Fisher M. Advising Autonomous Cars about the Rules of the Road. Electronic Proceedings in Theoretical Computer Science. 2022 September;371:62-76. Available from: http://arxiv.org/abs/2209.14035.

[11] Alves GV, Dennis L, Fisher M. A Double-Level Model Checking Approach for an Agent-Based Autonomous Vehicle and Road Junction Regulations. Journal of Sensor and Actuator Networks. 2021 September;10(3):41.

[12] Sartor G, Wyner A, Contissa G, Davila J. Mind the Gap - The Rules of the Road for Humans and Machines. In: Proceedings of the 18th International Workshop on Juris-Informatics (JURISIN 2024); 2024. p. 55-70.

[13] Leenes RE, Lucivero F. Laws on Robots, Laws by Robots, Laws in Robots: Regulating Robot Behaviour by Design [SSRN Scholarly Paper]. Law, Innovation and Technology. 2014 November;6(2)(2546759). Available from: https://papers.ssrn.com/abstract=2546759.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251630*

# Enhancing Automated Decision-Making in Administrative Law Through Digital Twins of Legislation

*A Grounded Theory Approach*

Florian SCHNITZHOFER [a] and Christoph G. SCHUETZ [a]
[a] *Institute of Business Informatics – Data & Knowledge Engineering, Johannes Kepler University Linz, Altenberger Str. 69, 4040 Linz*
ORCiD ID: Florian Schnitzhofer https://orcid.org/0009-0005-3338-0366, Christoph G. Schuetz https://orcid.org/0000-0002-0955-8647

**Abstract.** Automated decision-making in administrative law holds the promise of considerable efficiency gains, enhanced consistency, and increased transparency. This extended abstract distills a qualitative study based on semi-structured expert interviews that adopts a grounded theory approach to investigate the preconditions, constraints, and opportunities for embedding the concept of Digital Twin of Administrative Law (DTAL) into legislative and administrative processes while safeguarding the Rule of Law.

**Keywords.** Legal Ontologies, Semantic Interoperability, Ontological Alignment, Digital Twins, Automated Decision-Making, Rule of Law, Administrative Law

## 1. Introduction and Methodology

Public administrations worldwide are increasingly exploring automated decision-making (ADM) systems to improve the efficiency and consistency of administrative processes such as taxation, social benefit determinations, permitting, and enforcement. However, deploying sub-symbolic aritifical intelligence, e.g., models trained using machine-learning algorithms, in public decision-making raises significant concerns regarding the *Rule of Law*, i.e., the principle that governance should occur through publicly disclosed, consistent, and impartial laws. The legal stakes are high, as poorly designed algorithms can undermine due process, transparency, and equality before the law [3].

Today, agencies and implementers often re-code identical provisions, producing duplication, inconsistent outcomes, and fragile maintenance when parameters change [10]. A potential solution would be to publish *digital twins of legislation*: authoritative, machine-readable counterparts to legal texts that serve as a single source of truth for computations and explanations [6, 8, 7]. Previous work [13, 12, 14] outlines the vision of integrating legal ontologies with digital twin technology in administrative law, proposing the promulgation of legislation together with a machine-readable, logically-founded representation. Building on the broader concept of digital twins from engineering [5, 15],

this study investigates how digital twins could support ADM in administrative law that complies with the Rule of Law.

This extended abstract summarizes the results of a qualitative study (see [4] for further information on qualitative research), where five anonymized interviews with nine senior experts—spanning the European Parliament (policy, IT strategy, innovation), a national parliament (legislative counsel), academia, notarial practice, and public-sector ERP software development—were conducted (each 60–90 minutes long, either in English or German), transcribed, and coded (open → axial → selective) to theoretical saturation. Rigor measures included constant comparison, memoing, member checks, and an audit trail[1]

## 2. Discussion and Conclusions

By analyzing the experts' narratives, we derived key themes that shed light on both the limitations of automating law and the preconditions for success in that endeavor. Our findings identify four critical success factors for enabling ADM: (1) embracing only partial transformation, acknowledging that not all decisions in law can be automated; (2) achieving semantic standardization and ontological alignment; (3) Transforming legislation into high-quality ADM supporting structures; and (4) adapting drafting processes for legislation and governance mechanisms to accommodate ADM. The results indicate that partial automation is feasible, particularly in deterministic domains such as taxation and subsidies but legal frameworks require hybrid models that allow for human judgment and interpretative flexibility.

The findings support the conceptual design of the concept of a *Digital Twin for Administrative Law* (DTAL). Drawing on methods used in industry for digital twins [2] as well as on the concept of blockchain-based smart contracts [9, 1], this research posits a four-layer structure for a digital twin for administrative law:

1. **Statutory Text**: the natural-language law as traditionally published.
2. **Ontology**: a formal structure encapsulating the law's semantics.
3. **Configuration Model**: adjustable parameters which would include the typical axioms of the ontology (e.g., tax rates, exemption thresholds).
4. **Executable Logic**: the computational rules for enforcement (e.g., formulae, conditionals).

The findings of this study are based on a small European sample. Futurer work will broaden the scope with cross-jurisdictional studies. Future work will also investigate application of the DTAL concept in other domains and formalize digital-readiness criteria for legislative pipelines.

Grounded in empirical evidence, our study investigates how the DTAL concept supports explainable ADM that complies with the Rule of Law. These findings serve to develop a blueprint for code-based administration, establishing a foundation for trusted and citizen-centric digital government. Associated DTAL artifacts, including ontologies, executable rules, and step-by-step implementation guides, will be released online[2] under an open license.

[1] The study concept and coding tree are available online [11].

[2] https://lawdigitaltwin.com/

## References


[1] Jason Grant Allen. Wrapped and Stacked: 'Smart Contracts' and the Interaction of Natural and Formal Language. In *Smart Legal Contracts: Computable Law in Theory and Practice*. Oxford University Press, 2022. ISBN 9780192858467. .

[2] Stefan Boschert and Roland Rosen. Digital twin—the simulation aspect. In *Mechatronic futures: Challenges and solutions for mechatronic systems and their designers*, pages 59–74. Springer, 2016.

[3] Danielle Citron and Frank Pasquale. The scored society: Due process for automated predictions. *Washington Law Review*, 89:1–33, 03 2014.

[4] Juliet M Corbin and Anselm L Strauss. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. SAGE Publications, Los Angeles, fourth edition edition, 2015.

[5] Michael Grieves. Digital twin: manufacturing excellence through virtual factory replication. *White paper*, 1(2014):1–7, 2014.

[6] Anita Lamprecht. *Digital Twins of Law: Embracing Complexity*, pages 115–135. Springer Nature Switzerland, 05 2025. ISBN 978-3-031-78595-5. .

[7] Denis Merigoux, Nicolas Chataing, and Jonathan Protzenko. Catala: a programming language for the law. *Proc. ACM Program. Lang.*, 5(ICFP), August 2021. . URL https://doi.org/10.1145/3473582.

[8] Monica Palmirani, Guido Governatori, Antonino Rotolo, Said Tabet, Harold Boley, and Adrian Paschke. Legalruleml: Xml-based rules and norms. In Frank Olken, Monica Palmirani, and Davide Sottara, editors, *Rule-Based Modeling and Computing on the Semantic Web*, pages 298–312, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-24908-2.

[9] The Accord Project. Accord project concepts. https://docs.accordproject.org/docs/accordproject-concepts.html, 2024. Accessed on: 22 October 2025.

[10] Dag Wiese Schartum. *From Legal Sources to Programming Code: Automatic Individual Decisions in Public Administration and Computers under the Rule of Law*, page 301–336. Cambridge Law Handbooks. Cambridge University Press, 2020.

[11] Florian Schnitzhofer. Support material for the qualitative study on enhancing automated decision-making in adminstrative law through digital twins of legislation. https://doi.org/10.5281/zenodo.17093649.

[12] Florian Schnitzhofer. Integrating legal ontologies and digital twin technology: Enhancing rule-of-law-compliant adm in administrative law. In *Proceedings of the ESWC 2025 PhD Symposium*, Portorož, 2025.

[13] Florian Schnitzhofer. The self-driving state: Automated decision-making in modern governance. In *Proceedings of the International Symposium for Law and IT*, Vienna, 2025.

[14] Florian Schnitzhofer and Christoph Schütz. Towards a scalable architecture for legal-ontologies integrated into digital twins of administrative law. In *SEMANTICS 2025 Developers Workshop*, 2025.

[15] Fei Tao, Fangyuan Sui, Ang Liu, Qinglin Qi, Meng Zhang, Boyang Song, Zirong Guo, and Andrew Nee. Digital twin-driven product design framework. *International Journal of Production Research*, 57:1–19, 02 2018. .

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*

*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/FAIA251631*

# Modeling U.S. Supreme Court Briefs with Computational Argumentation

Heng ZHENG [a,1], Dexter WILLIAMS [b], Bertram LUDÄSCHER [c]
[a] *School of Information Science, University of Kentucky*
[b] *The Information School, University of Wisconsin-Madison*
[c] *School of Information Sciences, University of Illinois, Urbana-Champaign*

**Abstract.** We show how to represent U.S. Supreme Court briefs as a structured argument model that relies entirely on the briefs' existing structure.

**Keywords.** U.S. Supreme Court, large language model, computational argumentation

Mitigating the public's misunderstanding of how the U.S. Supreme Court makes decision is important [1]. Instead of predicting the outcomes of Supreme Court decisions, we show how to transform the legal arguments presented during the decision-making process into a more digestible form for the public. We focus on the arguments from the Supreme Court briefs, which can influence Supreme Court judges' opinions [2].

Supreme Court briefs, filed after a *writ of certiorari* is granted, can be modeled as a structured argumentation framework because they follow a strict structure [3]. This representation helps the public and stakeholders to understand what arguments were made on their behalf (by the petitioner, respondent, or *amicus curiae*), and how the debate unfolded.

We chose the *Acheson Hotels, LLC v. Laufer*[2] case as an example. Figure 1 shows the table of contents of three sample briefs and how we modeled them: 1) petitioner's brief that was filed before the respondent's brief; 2) the respondent's brief that was filed in response to the petitioner's brief; and 3) one amicus brief that supports the respondent.[3] We model each brief as a structured argument. The subsections under the ARGUMENT section are premises that support the conclusion (content in the CONCLUSION section). Here, we deem each subsection to support the conclusion independently, and the inference from the subsection, as a premise, to the conclusion is defeasible. For example, the part of the respondent's brief we showed in Figure 1 has Sections I and II that support the conclusion, and Section I is supported by Sections I.A, I.B, and I.C.

Attacks are from a later-filed brief to an earlier-filed brief, and the two briefs should not favor the same sides. By keeping the order in which attacks were made, we maintain the decision-making process. Otherwise, it would be counterintuitive for a party to file a brief to attack future arguments. Therefore, arguments in the respondent's brief and the amicus brief in Figure 1 attack the argument in the petitioner's brief. Attacks are classified as undermining (targeting a premise), undercutting (targeting the support

[1] Corresponding Author: Heng Zheng, hengzheng@uky.edu
[2] *Acheson Hotels, LLC v. Laufer*, No. 22–429 (U.S. Dec. 5, 2023).
[3] Petitioner's brief, Respondent's brief, and Amicus brief from *Disability Rights Education & Defense Fund*

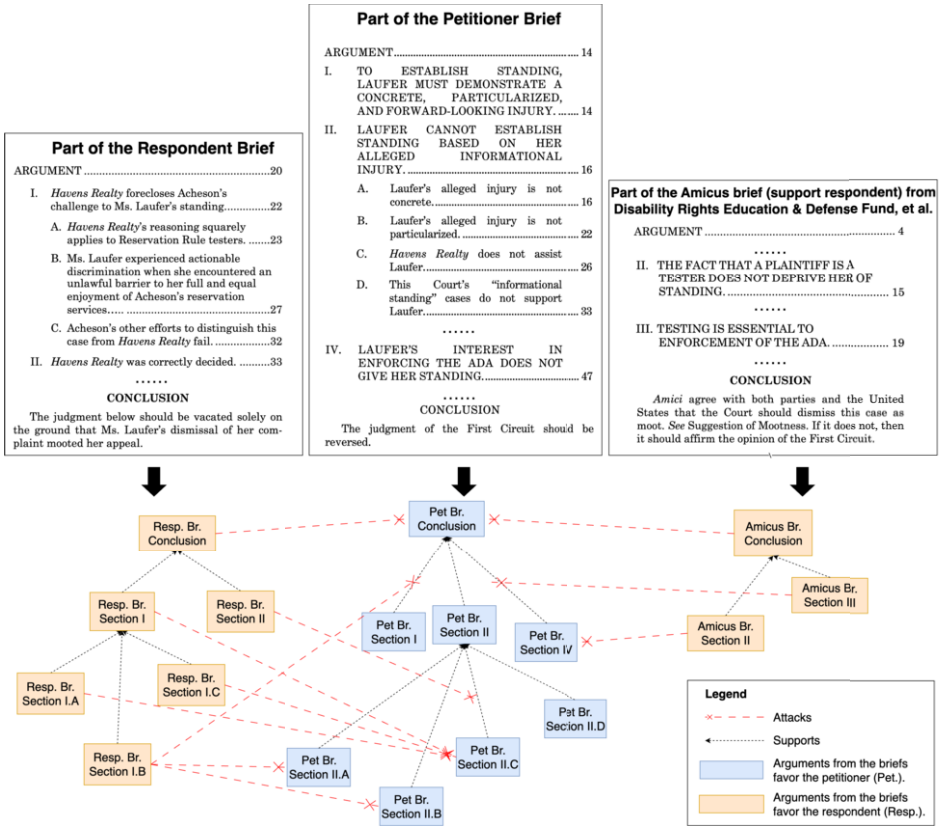



**Figure 1.** Modeling excerpts from the petitioner brief (middle), the respondent brief (left), and an amicus brief that supports the respondent (right) as a structured argument model

relationship), and rebutting (targeting a conclusion). For example, the respondent's brief concludes that the First Circuit's judgment should be "vacated," which is a rebuttal of the petitioner's brief's conclusion: the judgment should be "reversed."

Constructing the structured argument models can be time-consuming. We explored whether LLMs can assist the construction process by summarizing the sections in the briefs and identifying the attacks between the briefs. LLMs have shown their potential for argument summarization using a benchmarking dataset of U.S. Supreme Court briefs [4]. To see if LLMs can identify attacks between the arguments, we used `Mistral7B` for an initial experiment on five briefs for the *Acheson Hotel* case, and compared the results with Author 1's manual identification of attacks. We chose `Mistral7B` because it performs relation-based argument mining successfully on several non-legal datasets [5]. While a lack of annotated data makes meaningful evaluation challenging, we found the output is still promising.

Our initial attempts support further investigation on combining the strengths of LLMs and computational argumentation to help non-legal experts analyze U.S. Supreme Court briefs. We will investigate how to use LLMs to generate the argument models automatically and use formal semantics, such as ASPIC+ [6], to interpret the briefs further.

*Acknowledgments* This work was funded by the 2025 Cardiff University-University of Illinois System Joint Research and Innovation Seed Grants Program. We thank Athena Hatt for suggesting U.S. Supreme Court Briefs. We thank Jodi Schneider and Timothy McPhillips for providing valuable feedback on an earlier draft of this work.

## References

[1] Barry Sullivan and Ramon Feldbrin. The Supreme Court and the People: Communicating Decisions to the Public. *University of Pennsylvania Journal of Constitutional Law*, 24(1):1–92, 2022.

[2] Pamela C. Corley. The Supreme Court and Opinion Content: The Influence of Parties' Briefs. *Political Research Quarterly*, 61(3):468–478, September 2008.

[3] Dan Schweitzer. U.S. Supreme Court Brief Writing Style-Guide Developments. *Journal of Appellate Practice and Process*, 19(1):129–156, 2018.

[4] Jesse Woo, Fateme Hashemi Chaleshtori, Ana Marasovic, and Kenneth Marino. BriefMe: A Legal NLP Benchmark for Assisting with Legal Briefs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13139–13190, Vienna, Austria, July 2025. Association for Computational Linguistics.

[5] Deniz Gorur, Antonio Rago, and Francesca Toni. Can Large Language Models perform Relation-based Argument Mining? In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8518–8534, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

[6] Sanjay Modgil and Henry Prakken. The ASPIC+ Framework for Structured Argumentation: A Tutorial. *Argument & Computation*, 5(1):31–62, January 2014. Publisher: SAGE Publications.

*Legal Knowledge and Information Systems*
*R. Markovich et al. (Eds.)*
*© 2025 The Authors.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*

# Author Index

Alexander, C.S. 156
Ali, S. 264
Amantea, I.A. 427
Amara, F.Z. 37
Anim, J. 412
Araszkiewicz, M. 2
Ashley, K. 216, 239
Atkinson, K. 145
Audrito, D. 356
Bachňáková Rózenfeldová, L. 350
Bagnato, S. 86
Baldoni, M. 270
Baran, S. 300
Baroglio, C. 270
Benedetto, I. 378
Berghegger, C. 382
Berijanian, M. 156
Billi, M. 395
Borges, G. 371
Bourgne, G. 398
Brandizzi, L. 14
Bresciani, P.F. 26
Cagliero, L. 378
Carpentier, M. 392
Černý, J. 330
Chakravarthy, S.M. 98
Chen, Q. 385
Chen, X. 421
Cheng, R. 385
Chu, Q.H. 121
Ciabattoni, A. 133
Consiglio, E. 86
Contissa, G. 294
Corazza, M. 26, 251
D'Amato, C. 37
Dal Pont, T. 168, 389
Dar, H. 412
Darji, H. 276
de Martim, H. 282
de Melo, G. 276
Deshmukh, S. 288
di Caro, L. vi, 294
Didio, F. 37
Dik, J. 49
Dong, H. 61
Doutre, S. 392
Duffourc, M. 98
Dzehtsiarou, K. 145
Ecobichon, T. 392
Fanizzi, N. 37
Fedurcová, N. 300
Ferooz, F. 26
Ferrigno, B. 395
Fidelangeli, A. 294
Florimonte, F. 145
Francioso, D. 37
Fungwacharakorn, W. 398
Galassi, A. 294
Galli, F. 294
Galloni, F. 86
Garzo, G. 74
Golpayegani, D. 400
Gostojić, S. 424
Governatori, G. 427, 433
Gray, M. 133
Grundler, G. 294
Gultekin, S. 86
Gunawan, J. 98
Gurský, P. 300
Harašta, J. 324
Heckelmann, M. 276
Hladká, B. 330
Hoeben-Kuil, L. 98
Horvát, Š. 300
Huang, H.-C. 306
Huang, S.-C. 228
Hubner, J. 168
Hübner, J. 389
Hučková, R. 350
Idelberger, F. 204
Kamble, P. 288
Kanwal, S. 412
Kerr, A. 400
Knap, T. 415
Kołacz, M. 98
Kollnig, K. 98

Kong, Y. 312
Kratsch, C. 276
Kraus, I. 330
Kvapilíková, I. 330
Lagioia, F. 86
Lasek-Markey, M. 400
Lewis, D. 400
Liga, D. 110, 412
Lindsay, C. 121
Liu, C.-L. 306
Liu, J. 418, 421
Liu, Y. 385, 385
Longo, G. 26
Ludäscher, B. 439
Mailly, J.-G. 382, 392
Marcos, H. 318
Marković, M. 424
Markovich, R. vi
Matković, J. 424
Micalizio, R. 270
Mignone, R. 294
Mírovský, J. 330
Molinari, M. 427
Morello, Y. 133
Moudjari, L. 382
Mumford, J. 145
Nai, R. 356
Nanda, R. 318
Nguyen, H.N. 121
Nguyen, L.-M. 121
Nguyen, T.D. 121
Nguyen, T.M. 121
Novotná, T. 324, 330
Ohm, M. 204
Orth, T. 156
Oshio, K. 337
Palmirani, M. 26, 251
Palumbo, A. 74
Pang, T.F. 156
Perrussel, L. 382
Pham, N.A.T. 121
Philippe, C. 382
Pisano, G. 433
Pour, V. 430
Prakken, H. 344
Quaranta, M. 427
Rapp, A. vi
Robaldo, L. 412
Rodrigues, G. 14
Rossi Reich, M. 86
Rotolo, A. 385, 395, 433
Rover, A. 168, 389
Rubini, G. 37
Rudohradská, S. 350
Sabo, I. 168, 389
Salas-Jimenez, K. 382
Santin, P. 294
Sapienza, S. 26, 251
Sartor, G. 86, 433
Sasdelli, D. 180
Satoh, K. 312, 371, 398
Savelka, J. 98, 239
Schifanella, C. vi
Schnitzhofer, F. 436
Schuetz, C.G. 436
Schutz Veiga, J. 318
Shao, H.-L. 228
Shen, W. 385
Shi, B. 156
Sileno, G. 264
Sokol, P. 350
Spada, I. 356
Stanovský, A. 330
Steenhuis, Q. 192
Steffes, B. 180
Steging, C. 364
Sulis, E. 356
Szoplák, Z. 300
Tang, J. 418
Tarasconi, F. 378
Tedeschi, S. 270
Toni, F. 121
Torroni, P. 294
Trifiletti, V.S.M. 356
van der Torre, L. 61
van Dijck, G. 98
van Engers, T. 264
van Woerkom, W. 344
Varga, D. 300, 350
Vieira, M. 389
Vogel, D. 204
von Maltzan, S. 204
Westermann, H. 98, 318
Williams, D. 439
Wyner, A. 433
Xiong, Y. 312
Xu, H. 216
Younus, A. 400

Yousefi, Y. 395
Yu, L. 61, 110
Yu, Y.-c. 228
Zbiegień, T. 364
Zhang, L. 239
Zheng, H. 439
Zheng, S. 312, 385
Zilli, L. 251
Zin, M.M. 371