Towards Supporting Legal Argumentation with NLP: Is More Data Really All You Need?

Santosh T.Y.S.S.[1], **Kevin D. Ashley**[2], **Katie Atkinson**[3], **Matthias Grabmair**[1]
[1]Technical University of Munich, Germany
[2]University of Pittsburgh, USA
[3]University of Liverpool, UK

Abstract

Modeling legal reasoning and argumentation justifying decisions in cases has always been central to AI & Law, yet contemporary developments in legal NLP have increasingly focused on statistically classifying legal conclusions from text. While conceptually "simpler", these approaches often fall short in providing usable justifications connecting to appropriate legal concepts. This paper reviews both traditional symbolic works in AI & Law and recent advances in legal NLP, and distills possibilities of integrating expert-informed knowledge to strike a balance between scalability and explanation in symbolic vs. data-driven approaches. We identify open challenges and discuss the potential of modern NLP models and methods that integrate conceptual legal knowledge.

1 Introduction

Law has been an attractive domain for AI in both symbolic knowledge representation and statistical NLP. Both strands share the common goal of supporting legal practice through enhancing legal research, document analysis, drafting, and decision making. A focal question distinguishing them remains whether, and how, the process of legal reasoning [1] underlying all textual data shall be explicitly represented or left to opaque components, such as generative language models or neural classifiers.

In principle, legal reasoning resembles IF-THEN-like inference. Legal rules are established from sources (statutes, regulations, precedent, custom, etc.) and mandate that certain consequences follow if factual requirements are met in a specific situation. In reality, however, such logic-like inferences are interwoven with areas of ambiguity, vagueness, and human discretion (Urbina, 2002). At the same time, legal orders evolve over time, continuously refining and adjusting to a dynamic world. In knowledge engineering communities, legal reasoning is characterized as 'defeasible' (Carlos, 2001) rather than monotonic. Rules that are applicable on their face can be trumped by special exceptions, conflicting superior rules, or by distinguishing the precedent from which the rule derives. Thus legal decisions are subject to change, as they can be overturned on appeal. The evolving nature of law to align with shifting social values leads to different legal conclusions. When two parties are in conflict and desire two different resolutions, their argument will combine law and facts in a way that is beneficial to their respective goals - through adversarial discourse (Khairoulline, 2007). Legal argumentation can be seen as an exercise in competitive theory formation in front of an arbitrator, with each side constructing arguments supported by evidence, written law, cases and other authority to favor their desired conclusions while addressing pitfalls of opposing theories (Rissland et al., 2003).

AI & Law as a field started started in the 1970s, when Buchanan and Headrick (1970) suggested that computer modeling of legal reasoning would be a promising area for research to better understand legal reasoning and argumentation. Many approaches have been proposed over the past three decades capturing several types of reasoning by means of symbolic representations. Some 50 years after the field's beginnings, the legal profession is experiencing considerable disruption by NLP technology, most prominently large language models (LLMs). In this paper, we provide a review of AI & Law work offering faithful modeling of legal reasoning but also requiring expensive legal expertise. We contrast this to modern, largely non-explainable, data-driven methods, which predict

[1]By 'legal reasoning', we refer to the wide range of activities involving interpreting, arguing, and applying legal principles to reach conclusions. Legal reasoning is not a single task but a collection of related tasks around the main theme of legal decision-making as the interrelation of more of less well-defined rules and societal values with the facts of a specific case towards an outcome. Given the limited space available, we use 'legal reasoning' as an umbrella term to cover the diverse contributions on this topic in the literature.

Proceedings of the Natural Legal Language Processing Workshop 2024, pages 404–421
November 16, 2024

legal conclusions directly without engaging in any explicit legal reasoning.

Our main contributions are as follows:

- An introduction to legal systems to sensitize readers to assumptions made in technical work
- Surveys of (1) landmark AI & Law work and its lessons learned, and (2) data-driven approaches to legal AI and legal NLP
- A detailed discussion of perspectives to unify both strands to meet future challenges.

Our discussion makes the following arguments:

- Future work on legal AI must strive to integrate legal expertise with data-derived models.
- Conveniently available legal NLP datasets come with structural assumptions, noise, and biases, which must be accounted for.
- Change of legal systems over time remains an under-explored aspect in NLP works.
- LLMs help alleviate knowledge acquisition bottleneck for domain model construction.
- There is value in NLP that produces and assesses arguments about legal conclusions in an explainable way with domain knowledge representation.
- Qualified evaluation in legal NLP is underdeveloped given the often non-well-defined nature of legal practice support tasks, resulting in exaggerated attention on convenient but uninformative benchmark metrics.

While prior surveys of Katz et al. (2023b) and Zhong et al. (2020b) focus on cataloging various use cases, tasks, and NLP techniques in legal AI, our paper critically examines the historical integration of expert knowledge into legal systems and advocates for its revival and synthesis with data-driven methods. We emphasize the unique value of expert-informed knowledge in ensuring legal reasoning aligns with established principles, which is not the primary focus of the aforementioned reviews. In contrast to Mahari et al. (2023), which highlights the disconnect between the tasks that are pursued in legal NLP research and the actual needs of legal practitioners, our work emphasizes the critical importance of integrating expert-informed knowledge to avoid this gap. We also present directions for synthesizing expert knowledge with current technological advancements, thereby overcoming traditional bottlenecks in knowledge acquisition and enhancing the efficacy of structured argumentation models.

Most importantly, we contribute a comprehensive distillation of the conceptual ideas developed and researched by the AI & Law community prior to the recently surging interest in law as an application domain for mainstream NLP. In part, our motivation is to connect these communities. Much legal NLP work does not build on formal models of legal knowledge and reasoning but characterizes it mostly as precursor work to modern statistical methods. Our position is that this view does not do justice to the insights gained and legal authenticity captured in this body of research. Symbolic AI & Law has thought about how to incorporate legal expertise in models more deeply than most current NLP works, and hence the fields should merge and learn from one another. We strive to drive home the necessity of a paradigm shift in legal NLP, one that values and integrates the profound expertise of domain specialists with the capabilities of data-driven technologies.

2 Legal Systems in a Nutshell

Legal systems revolve around legal subjects, institutions and actors, and sources of law. While there is variation across settings, the most relevant sources typically comprise a national constitution, primary legislation (often referred to as ‘statutes’, etc.), secondary ‘executive’ regulation, precedents decided by courts, and other auxiliary sources. A major division exists with regard to the role of precedents relative to written law, as well as the methodology of arguing with them. Legal systems primarily influenced by continental Europe follow the ‘civil law’ approach, where important decisions are mostly condensed into context-free interpretive rules to codified law that are compiled in secondary literature (e.g., so-called ‘commentaries’). In parts of the world with primarily English legal influence, so-called ‘common law’ systems, precedents are regularly applied by means of analogizing and distinguishing arguments that take into account the facts of the case in much greater extent than civil-law-type reasoning will. International courts (e.g., the European Court of Justice, the European Court of Human Rights) usually follow hybrid methodologies that are specific to the legal regime they govern. Despite some recent diversification, virtually all AI & Law research comes from either civil or common law backgrounds and makes corresponding assumptions, which is why we include this introduction. It is important to note that this coarse systematization is a great simplification of the world's diverse legal systems and cultures, and only intended to supplement our survey.

Written law generally consists of primary (i.e., parliamentary) legislation and secondary (i.e., executive) regulation. While enacted by two different branches of government, they are structurally similar in that they encode rules that can be formalized in IF-THEN relations. They also contain ambiguous and vague formulations in need of interpretation, for which special methods exist that are beyond the scope of this work. It is typically up to the judiciary (i.e., the courts) to settle open questions through landmark cases, often after arguments being developed in academic literature. These decisions then become part of the discourse accordingly in the applicable methodology. This transition from rule-based to case-based reasoning (in the common law) has intuitively been termed "when the rules run out" by (Gardner, 1987).

When arguing a case relative to a precedent, it is a fundamental principle of justice that similar cases should be treated alike. In common law jurisdictions, this principle is formalised in the doctrine of *stare decisis*, which obliges decisions of the appropriate status to be followed when deciding a new case. Civil law legal orders also recognize a binding effect of high court precedent, but argue with them differently. While higher court decisions bind lower courts, cases move in the opposite direction. They are first filed in, for example, district or trial courts, where evidence is heard and first decisions are made. Decisions can then be appealed to the Appeals Courts, and eventually to Supreme Courts. At some point in this progression, arguments on evidence will be considered settled and only purely legal errors will be permissible grounds for further escalation. In a legal system, such 'appeals tracks' exist for various jurisdictions (civil, criminal, administrative, etc.) and can be spread across geographic entities (e.g., federal vs state courts).

It is worthwhile to acknowledge that legal systems are inherently human-centric, involving complex decision-making processes where persuasion, interpretation, and subjective judgment play critical roles. Legal decisions are not solely about determining which side should win as a matter of justice, but about who can present the most convincing argument within the framework of established laws, principles, and precedents. The main vision of AI & Law is that state and private actors in all aspects of the legal system can benefit from supporting software that seamlessly connects to the concepts and concerns they have been trained for and work with. Notably, legal reasoning not only happens in courts, but also in public administration and law enforcement (i.e., the executive branch), where law needs to be applied to specific situations (e.g., permits, taxes, public safety, etc.). Human accountability is paramount for the trust in the overall workings of a democratically governed society. Hence, this vision is one of AI supporting human decision makers and not replacing or unduly influencing it.

3 Knowledge-based Approaches

AI & Law research started with modeling of legal reasoning by means of knowledge representation.

Rule-based Approaches Early landmark work demonstrated how British immigration law could be represented in Prolog (Sergot et al., 1986) and outlined challenges faced in this process, including the law's rule-exception pattern, negation-as-failure (i.e., failure to prove true) vs. classical negation (logical, certain falseness), and counterfactual reasoning. Waterman and Peterson (1980) developed a specialized language for rule-based legal inference. Rules establish conclusions from antecedents in a forward/backward chaining manner, thereby spanning open a derivation tree of a case outcome. They justify a position and explain how a conclusion can be reached, but they do not capture the dialectical aspects associated with argumentation, since no conflicting arguments are generated and no indeterminacy is accounted for. Gardner (1987) extended by using augmented transition networks to model contract formation over time given agent actions with a basic form of uncertainty - If a condition was a 'hard question' and could not be decided, the network would fork into two alternative ways to legally treat the facts. Overall, early rule-based systems were still predominantly derivations rather than argumentation models, although they correspond well to how lawyers analyze cases.

Case-based Approaches The adversarial nature of law naturally demands to represent arguments for both dispute sides. The precedent-focused nature of US common law was a suitable domain for the development of what became known as 'legal case based reasoning' systems. In the prominent TAXMAN system, McCarty (1976) modeled the majority's and minority's theories and arguments in the famous tax law case of *Eisner v Macomber*, *252 U.S. 189 (1920)* (Eisner v. Macomber), surveying the intricacies one must account for, if resolved to capture all decision-relevant concerns in depth. The HYPO system (Ashley, 1991) modeled parts of US Trade Secrets Law by means of dimensions. These are typical fact patterns that favor different sides of the dispute, and can be used to analogize and distinguish cases argumentatively by means of set comparison. The focal concept here is a 'three-ply argument': A proponent cites the most-on-point precedent with the greatest factor overlap. The opponent distinguishes by pointing to a disfavorable factor in the precedent but not in the new case, or a favorable factor in the current case but not in the precedent, and cites a counterexample precedent. Finally, the proponent offers a rebuttal by distinguishing the counterexample. This was built upon in the CATO system (Aleven, 1997), which arranged 'factors' into a hierarchy, on the basis of which more sophisticated argumentation was possible (e.g., using hierarchy parent factors).

Hybrid & Extended Systems CABARET (Rissland and Skalak, 1991) first combined rule-based reasoning with HYPO-style case based reasoning around ill-defined terms contained in the rules. The integration is performed via a collection of control heuristics that interleave arguments of both kinds to support a particular conclusion. GREBE (Branting, 1991) further extends this hybrid architecture with formalized domain knowledge and a semantic network representation to retrieve and compare cases. BankXX (Rissland et al., 1996, 1997) embeds HYPO-style factor-based reasoning with a domain model into a 'legal theory space' that can be searched for plausible arguments.

Integration with Prediction CATO had been developed as a tutoring system and did not predict case outcomes. Issue-Based Prediction (IBP) (Bruninghaus and Ashley, 2003) extended the factor-based representation with a model of legal 'issues', each of which could be predicted via case-based reasoning. Ashley and Brüninghaus (2009) even proposed SMILE + IBP, classifying the presence of factors in cases by means of NLP, whereas prior factor-based systems had all relied on manual factor coding of cases. It pioneered data-driven approaches for ascribing factors to be used in conjunction with a domain model without circumventing the reasoning process entirely.

Values, Time, and Procedure: Berman & Hafner explored deeper aspects of representing cases, many of which remain challenging to this day. Berman and Hafner 1993 proposed to supplement each factor with "legal purpose(s) which it affects, and each legal purpose in turn specifies whether it favours the plaintiff or defendant". Parties may offer competing arguments based on factor-based case analogies. *Teleological* knowledge allows a model to go beyond factual similarities to include broader jurisprudential concepts. This was highly influential in subsequent work (Greenwood et al., 2003; Chorley and Bench-Capon, 2005b; Wyner et al., 2007; Grabmair and Ashley, 2011; Muthuri et al., 2017; Grabmair, 2017; Maranhao et al., 2021), which converged towards speaking of "values" rather than purposes.

Berman and Hafner (1995) contributed a pioneering model of the temporal dynamics of case-based legal reasoning: "legal precedents are embedded in a temporal context of evolving legal doctrine, which can result in a strong precedent becoming weaker over time, to the point where a skillful attorney could reasonably predict that it will no longer be followed." This temporal dimension has also received attention in other works (Rissland and Xu, 2011; Henderson and Bench-Capon, 2019; Prakken and Sartor, 1998; Branting, 1993).

Berman and Hafner 1991 observe that the support of a precedent decision for a case to be argued is linked to its respective procedural setting. They distinguish the pleading, pre-verdict, and verdict stage. A further difference exists between decisions on procedural matters and decisions on matters of fact and/or law. A decision in favour of the defendant party based on a procedural matter (e.g., lack of evidence) may not support the same decision in a new case which shares the factual features of the precedent but is to be decided on its merits. The question of decision context has received limited attention in subsequent works (e.g., Wyner and Bench-Capon 2009; Verheij 2016). Even in the recent works on NLP-based legal judgment prediction, case outcomes are often greatly simplified, up to the point of an impoverished binary variable of whether a party won the case or not.

Theory Construction Approach: As McCarty 1995 pointed out, "[T]he task for a lawyer or a judge in a hard case is to construct a theory of the disputed rules that produces the desired legal result, and then to persuade the relevant audience that this theory is preferable to any theories offered by an opponent". Bench-Capon and Sartor (2000, 2003, 2001) model a 'theory' as a set of factor-based rules and preferences among them derived from value preferences. Different theories can be compared with reference to the number of cases whose outcome they explain. The rules and preference relations form tradeoffs between sets of values raised by factors in the cases. These establish preferences among rules, which in turn predict case outcomes. The CATE system (Chorley and Bench-Capon, 2004) enabled manual creation and testing of theories as prolog programs. The AGATHA system (Chorley and Bench-Capon, 2005a) constructed theories autonomously using A* search.

Computational Argumentation: Producing an argument by using a rule-driven strategy implemented with case-driven argument moves remains a central way of justifying conclusions in cases. In the 90s this was mainly pursued using dialogue games which were designed to allow an adversarial discussion between the two parties, one represented by the computer and one by the user. Examples include Gordon 1993; Hage et al. 1993; Prakken and Sartor 1997, 1998; Loui and Norman 1995. While many of the systems referenced thus far model argumentation ad hoc, the AI & Law field interacted considerably with its neighboring discipline of general computational models of argumentation. Of particular interest in this context is the concept of 'argument schemes' as well as the connection to models of so-called 'abstract argumentation'.

Argument Schemes: An argument scheme is a stereotypical pattern of reasoning primarily constituting a claim, a set of positive premises, and, optionally, a set of negative exceptions. Argument schemes have a long history, as laid out in (Macagno et al., 2017). In modern times, schemes were used by Perelman and Olbrechts-Tyteca 1969 and Toulmin 1958. In AI & Law, the Toulmin argument model had been historically popular. It recognizes different roles of statements in an argument: Claim, Qualifier/Strength, Data/Premises, Warrant/Inference, Backing, and Rebuttal. This is suitable for legal reasoning by incorporating authority for the warrant and by including a rebuttal component in recognition of the defeasible nature of legal reasoning. Walton (1996) introduced a variety of schemes into AI & Law (e.g., from Expert Opinion, from Negative Consequences, from Rules, etc). Verheij (2001); Gordon and Walton (2009) supplemented them further (e.g., from position to know, from ontology, from cases, from testimonial evidence). Schemes have become central in AI & Law research, being used in reasoning with evidence (Bex et al., 2003; Bex, 2011), reasoning with cases (Prakken et al., 2015), e-democracy (Atkinson et al., 2006), statutory interpretation (Araszkiewicz, 2021), and value-based argumentation (Grabmair, 2016; Greenwood et al., 2003).

Abstract Argumentation Framework: In seminal work, Dung (1995) defined *abstract argumentation frameworks* (AAFs), which were introduced to AI & Law by Prakken 1995. An abstract argumentation framework comprises a set of arguments and set of attack relations between them. The justified arguments are then evaluated based on subsets of arguments ('extensions') defined under a range of semantics. The abstract nature of Dung's theory says nothing about the structure of arguments, the nature of attack or defeat, or use of preferences. This opacity, and the coupling with argument schemes, motivated the development of structured argument models. For example, ASPIC (Caminada and Amgoud, 2007) adopts an intermediate level of abstraction by making some minimal assumptions on the nature of the logical language and the inference rules, and then providing abstract accounts of the structure of arguments, the nature of attack, and the use of preferences. Prakken 2010; Modgil 2009 generalised the ASPIC framework to develop ASPIC+, which can capture a broader range of systems with various assumption-based argumentation and systems using argument schemes. ASPIC+ has been applied to study legal reasoning in the works of Prakken 2012; Prakken et al. 2015.

Abstract Dialectical Frameworks (ADFs) (Brewka et al., 2013) generalize the AAF representation to node-and-directed-relations form with a set of local acceptance conditions. This allows both attack and support influence, resulting in an abstract yet intuitive model for legal reasoning. For example, the ANGELIC method (Al-Abdulkarim et al., 2016b) uses ADFs for representing case law in an explainable inference model on the basis of a hierarchical factor representation. The maintainability of such a representation is discussed in (Al-Abdulkarim et al., 2016a). Atkinson et al. (2019) extended to reasoning about factors with magnitude, thereby going beyond purely boolean proposition representations of cases.

Overall, the advantages of knowledge-based approaches are that they explicitly model legal reasoning and provide explanations of inferences.

4 Data-driven Approaches

Relationship to Political Science Research: Data originating in the legal system has been the subject of extensive analytical study in the field of *empirical legal studies*, including court and judge decision/voting behavior (e.g., Segal 1984; Kort 1957; Nagel 1963; Ruger et al. 2004). As most of them neither model legal reasoning nor apply NLP techniques, we do not include them in our survey.

Early AI & Law: Knowledge-centered approaches can achieve high degrees of faithfulness in their representation and explainability in their inferences, but face the 'knowledge acquisition bottleneck', as they require large amounts of expertise and modeling effort. This is in contrast to data-driven models with less hand-crafted expertise. Early works by Mackaay and Robillard (1974) used nearest-neighbor methods for outcome classification. In the 1990's, Pannu 1995; Bochereau et al. 1991; Philipps 1989; Bench-Capon 1993 trained neural networks to predict outcomes and derive input feature weights. Unsurprisingly, such early applications of ML attracted criticism (Aikenhead, 1996; Hunter, 1994). Obtaining substantial amounts of processable data was challenging and extensive feature engineering was necessary. These works focused on the application of neural networks to identify how influential certain information is for the decision and did not engage in comparative benchmarking.

Towards Modern Legal NLP: Recent years saw a resurging interest in case prediction through the use of data-driven methods learning from the large datasets now available from different jurisdictions, such as the ECtHR (Chalkidis et al., 2019, 2022a, 2021; Aletras et al., 2016; Medvedeva et al., 2021; SAYS, 2020; Tyss et al., 2023b,a; Santosh et al., 2024c; Liu and Chen, 2017; Medvedeva et al., 2020; SAYS, 2020) Chinese Criminal Courts (Luo et al., 2017; Yue et al., 2021; Zhong et al., 2020a, 2018; Yang et al., 2019), , US Supreme Court (Katz et al., 2017; Kaufman et al., 2019), Indian Courts (Malik et al., 2021; Shaikh et al., 2020) French court of Cassation (Şulea et al., 2017b,a; Bertalan and Ruiz, 2020) Supreme Court of Switzerland (Niklaus et al., 2021), Turkish Constitutional court (Sert et al., 2021), UK courts (Strickson and De La Iglesia, 2020), German courts (Waltl et al., 2017), Brazilian courts (Lage-Freitas et al., 2022) and Philippine courts (Virtucio et al., 2018).

Earlier works employed bag-of-words features (Aletras et al., 2016; Şulea et al., 2017a,b; Virtucio et al., 2018; Shaikh et al., 2020; Medvedeva et al., 2020). More recent approaches use deep learning techniques (Zhong et al., 2018, 2020a; Yang et al., 2019) involving convolutional or recurrent networks followed by adoption of pre-trained transformer models (Chalkidis et al., 2019; Niklaus et al., 2021), including legal-domain specific pre-trained variants (Zheng et al., 2021; Chalkidis et al., 2020, 2023; Douka et al., 2021; Masala et al., 2021; Xiao et al., 2021; Hwang et al., 2022; Niklaus et al., 2023). Classification tasks on legal text interrelate, and so other words have leveraged dependencies between tasks for improving models (Santosh et al., 2023a; Yue et al., 2021; Valvoda et al., 2023; Zhong et al., 2018; Feng et al., 2022; Ma et al., 2021; Dong and Niu, 2021; Yang et al., 2019; Huang et al., 2021; Hu et al., 2018) and added additional loss constraints (such as contrastive learning exploiting label information), (Tyss et al., 2023b; Zhang et al., 2023; Gan et al., 2022; Liu et al., 2022) and injected legal knowledge (Liu et al., 2023; Santosh et al., 2023b, 2024c; Gan et al., 2021; Zhong et al., 2020a; Feng et al., 2022)

Overall, one can observe a trend towards applying NLP models to legal text with little to no architectural bias or explicit domain representation. These are then compared along quantitative metrics, typically with regard to high level classification/prediction goals (e.g., case outcome variables and document-level keywords) at the cost of interpretability. As Berman & Hafner have observed in the 1990s, however, case outcomes are highly contextual in time, procedure, and socio-legal purpose. Classification benchmarks risk decoupling a sense of technical progress towards a notion of model 'understanding' from supporting a realistic task (e.g., legal argumentation) by focusing on a highly reductive representation of its outcome. For instance, case outcome predictions are often treated as binary targets based on the majority opinion, even though judges on the same bench frequently have conflicting reasoning, leading to dissenting or concurrent opinions (Xu et al., 2024). This reductive approach overlooks the nuanced legal argumentation underpinning each decision, focusing on a single outcome instead of capturing the depth of legal reasoning and debate.

Limits of Classification Benchmarks: The working assumption of these approaches is that by getting better at the benchmark, models encode more legal knowledge which can be extracted as explanations for predictions. To the best of our understanding, however, this promise has not been fulfilled. Initial works on data from the EtCHR, Aletras et al. 2016; Chalkidis et al. 2019 listed words based on feature importance or highlighted text based on attention scores. In later works, Chalkidis et al. (2021) used regularization techniques to identify paragraphs that support a finding of a violation of ECtHR. The extracted rationales did not correspond well to the annotation by a single legal expert. Santosh et al. (2022); Malik et al. (2021) continued the trend of computing paragraph level importance using interpretability techniques such as Integrated Gradient and tried to assess them against expert-annotated important paragraphs, also with only moderate success. In the ECtHR context, Santosh et al. (2022) discovered evidence that BERT-based classifiers rely on shallow predictors. This can be mitigated using adversarial training, but alignment still remains low. Recently, Xu et al. (2023) assessed rationale alignment at the more difficult, fine-grained word level. The experiment uncovered inconsistencies in the court metadata and illustrated how even annotations by two legal experts may not align well. To add to the challenge, a pilot study by Branting et al. (2021) discovered that human performance in a prediction task does not improve if users are given access to a saliency map derived from a prediction model. Recent work by Mumford et al. (2023b) reported that human performance on the judgment prediction task closely resembled randomness and was unaffected by domain knowledge. These results all cast doubt on the assumption that, at least for classifiers models, benchmark performance correlates with better explanations. The data may be noisy, the labeling too simplified, the predictors too shallow, the expert disagreement low, and the utility of a salience map limited. It should also be noted that the potential leakage of benchmark test data into training corpora remains under-discussed and unmeasured.

Other body of works on outcome classification of ECtHR cases predict the decision from a textual description of the case facts alone. By contrast, what lawyers actually need is the explanation why the resolution of a case is the proper application of the law and in line with what traditional AI & Law work would call a 'theory' of ECtHR jurisprudence. The outcome must be based on a justification which presents equitable arguments, can be reviewed on appeal, and hold up under public scrutiny.

Shift to Generative Models: LLMs have also been evaluated against case outcome classification as a benchmark. Chalkidis (2023); Vats et al. (2023); Trautmann et al. (2022); Shui et al. (2023) tested various early models and found them to score relatively low in quantitative metrics, which stands in contrast to their scores on some bar exams (Katz et al., 2023a; Freitas and Gomes, 2023). They report on experiments with several models and prompting techniques, including zero/few-shot prompting, prompt ensembling, chain-of-thought, and activation fine-tuning. Yu et al. (2022, 2023) employ prompts that are derived from legal reasoning methods (such as the common law IRAC (Issue, Rule, Application, Conclusion). Trautmann (2023) uses prompt chaining with an initial summarization step to deal with lengthy legal documents. Jiang and Yang (2023); Deng et al. (2023) develop syllogism prompting providing the three deductive reasoning steps for major premise (article/law retrieval), minor premise (element extraction from facts) and conclusion (judgement).

LegalBench (Guha et al., 2023) recently presented the first aggregated benchmark beyond classification-like evaluation to test the reasoning abilities of generative models. Kang et al. (2023) applies the IRAC methodology comprehensively to LegalBench subtasks. While ancillary challenges remain (e.g., the need to manually assess model performance non certain tasks), this development is in line with our arguments in this paper.

5 Challenges & Future Directions

Combining Knowledge and Data: The pressing question is how best to integrate legal knowledge and ML so that a system can learn from data and still seamlessly interface to a lawyer's understanding of the domain by means of a conceptual representation. A number of such hybrid systems can be found outside of NLP: *Split Up* (Stranieri et al., 1999) combined expert-crafted rules and neural networks trained from data in a factor-based model of Australian family law to predict divorce asset division. In the CATO line of work, both AGATHA (Chorley and Bench-Capon, 2005a) and VJAP (Grabmair, 2017) leveraged structured legal argumentation for prediction with signals derived from a case base. Moving to NLP, one intuitive combination is to ascribe factors from cases using text processing and proceed with formalized legal inference. This was employed in SCALE (Branting et al., 2021) to enable a logic model to predict WIPO domain name disputes, and in the ECtHR domain by inference using an ADF representation Mumford et al. (2022, 2023a). Gray et al. (2023) automatically identified factors in Fourth Amendment auto stop cases, demonstrated their predictive value, and used ML techniques to explain case outcomes in terms legal professionals can understand. Holzenberger and Van Durme (2021) apply neural models to identify argument slots in legal provisions and find suitable filling elements from fact descriptions, thereby enabling rule-based inference. Similarly, Holzenberger and Van Durme (2023) automates the translation of cases into a knowledge base by posing it as an information extraction task.

Data Utilized: Ideally, outcome prediction systems in the legal domain should rely on the information available before proceedings start and legal conclusion are determined (e.g., argumentative memoranda from the parties). Most case outcome classification research is conducted based on fact descriptions that are taken from judgments. These are often highly selective summaries tailored to align with the decision (Tippett et al., 2021). Although they may not explicitly contain outcomes, this can introduce confounding effects as demonstrated in Santosh et al. (2022). To illustrate the effect of proxy data on performance, Medvedeva et al. (2021) utilized data from ECtHR 'communicated cases', court-prepared summary data derived from applicant submissions, published before trial and observed a decline compared to facts statements from judgments, highlighting the need to select appropriate data for this task to draw reliable conclusions (Medvedeva et al., 2023; Medvedeva and Mcbride, 2023). Such work may also be subject to data selection bias related to which cases reach which court, and with regard to how they are published. For example, a higher court will receive a different distribution of cases (i.e., such with grounds for appeal) than a district court, and only a subset of them may be published. Finally, many cases are settled before or during trial, further skewing the dataset (Osbeck and Gilliland, 2018).

Temporal Dynamics Current legal NLP methods often operate under the implicit assumption that past training data is homogeneous and neglect its sequential nature. In reality, attitudes and case law change over time, with later cases altering and superseding the roles of older ones. All shifts in jurisprudence confront the model with a cold start problem of little training data for a new legal rule and copious training data for outdated ones. These dynamics can in principle be modeled. For instance, overruling detection can identify where previous legal precedents have been overturned, and trigger techniques such as model unlearning (see Nguyen et al. 2022) or selective forgetting. One can also strive to detect updates in beliefs/knowledge expressed in decisions over time, and modify such beliefs within the model (Hase et al., 2021). Santosh et al. (2024b) accounts for the temporally evolving nature of classification tasks on legal data using continual learning approaches.Overall, however, the temporal dynamics of legal corpora remain largely unaddressed in recent works.

Domain Model Construction Rule-based models of the law are powerful tools to develop software that supports legal practice, but constructing them demands considerable legal expertise. Modern LLMs put us into a position to create these structures in a (semi-) automated fashion. Savelka et al. (2023) shows constructive evidence of this, but it remains an open questions whether LLMs can systematize large complexes of legal source material into well-formed, legally correct representations. Ascribing factors from facts text in unseen cases by means of developing classifiers requires training data relative to an exhaustively defined list of factors. The more likely scenario is that generative models can be prompted with specific facts to subsume them under a factor pattern description. For example, Gray et al. (2024) applied generative AI automatically to identify factors in Fourth Amendment auto stop cases.

From Argument Mining to Generation: The task of constructing abstract argumentation models closely dovetails with the field of argument mining (i.e., the detection of argumentative text segments and their interlinking). Traditionally, argument mining mainly encompasses four sub-tasks as formalized by seminal work in Palau and Moens 2009: text segmentation, argument span detection, classification (e.g., conclusion, premise), and prediction of graph relations between spans. Follow up work by Wyner et al. 2010; Grabmair et al. 2015; Poudyal et al. 2020; Habernal et al. 2023; Ali et al. 2022, 2023; Grundler et al. 2022 focused on the first three subtasks, with fewer models engaging in graph construction. Modeling the relationships and comparative strength between conflicting arguments is a crucial piece to connect these extractive argumentative mining efforts to structured argumentation, largely unaddressed by existing works.

Even with powerful LLMs available, optimal argumentation support systems for legal practitioners benefit from structured representations of legal information and argumentation. While argumentative text can now be generated by current models, it remains a challenging cognitive task to systematize and assess arguments strategically. A productive support system should produce arguments in a transparent manner, and offer the user an intuitive way of resolving multiple complex arguments towards a justification of a decision. Naturally, this also entails questions around mindful interface design and organizational processes to facilitate accountable human decision making where capable text generation systems are accessible.

Role of Evaluation: The true value in NLP for legal applications lies in producing, structuring, and assessing arguments about legal conclusions in an explainable way so that they may maximally support human experts. This human-centric nature of legal systems introduces a level of complexity that purely data-driven systems often struggle to capture when classifying variables from close-to-raw data. By the same token, LLMs may generate text that may seem lawyer-like, but integrating them in processes of legal practice regularly involves interfacing them with symbolic data structures on both input and output ends, as well as maximizing consistency and correctness of generated text in ways that is defined by the legal concepts of the application context. This may include obfuscating cumbersome and error-prone model prompting behind traditional user interfaces composed of elements that map to symbols in the domain (e.g., types of contract clauses, factor-like aspects of cases, information elements of interest to draft process memoranda, etc.). The complexity of human legal decision-making highlights the inadequacy of current evaluation metrics. Legal NLP works should, ideally, tangibly indicate progress towards optimal argumentation support systems for legal practitioners, yet frequently convenient evaluations are prioritized over informative ones. This is, of course, due to the nuanced and often ill-defined characteristic of legal practice tasks. Still, legal databases are more than large repositories of text for autoregressive pre-training, but resources for tackling these use cases, including, for example, using prior decisions in constructing and responding to arguments. Legal NLP's efforts should be evaluated - and reviewed - in terms of how well models provide such functionality (Ashley, 2022). Many legal NLP works specify use cases, yet few account for them in their evaluative framework by conducting studies with legal experts, or benchmark their automatic metrics against human evaluations. Research on evaluation criteria that better capture the practical utility of legal NLP systems in real-world settings should be among our top priorities.

Examples of human evaluations in specified use cases include the following: In Elaraby et al. (2024) human experts evaluated the legal argument coverage in generated summaries. In Mullick et al. (2022) and Salaün et al. (2022), humans assessed legal summaries' relevance, readability, fluency, or adequacy. In Xu and Ashley (2022) expert evaluators assessed the information quality of legal summaries in terms of generated question-answer pairs. Experts evaluated the legal importance of automatically identified paragraphs in Santosh et al. (2022) but achieving expert annotation agreements is challenging, especially given noisy metadata (Xu et al., 2023). Evaluations benchmarked human classification of case verdicts under ECHR Article 6 in Mumford et al. (2023b) and compared expert annotations to automatically generated explanations in Malik et al. (2021) and to automatically identified factor sentences in Gray et al. (2023).

6 Conclusions

We believe that knowledge-based approaches to building legal argument support systems deserve the attention of the modern NLP community, as they embody a culture and method of capturing intricacies of legal systems and argumentation that are often simplified away in the increasingly easier application of large mainstream models to legal data. The prominent role of benchmarks compounds this by drawing attention towards quantitative progress instead of real, empirical investigations of downstream benefit to practitioners. At the same time, LLMs widen the knowledge acquisition bottleneck for structured models considerably, opening up new opportunities. We believe there is great value in combining knowledge- and data-driven systems rather than continuing the assumption that deep expertise will reliably emerge given large enough amounts of data and computation.

Limitations

This paper focuses on legal NLP as applied to tasks that involve the application of legal source material to case facts, analysis of case texts, and legal argumentation in general. Other subfields of NLP in the legal domain do not focus on argumentation about the legal significance of case facts, such as technology-assisted review in e-Discovery, contract analysis, and patent search. Similarly, legal question answering, automatic summarization of judgments, legal information retrieval, and models supporting regulatory compliance, although important, are in focus for our argumentation-related narrative. We strive to synthesize a very broad notion of the important role of expert legal knowledge to facilitate better NLP systems that will be of high utility to the stakeholders involved in the ecosystem. In our way, the way forward requires input from diverse perspectives and collaboration across multiple disciplines, including law, computer science, linguistics, and ethics to achieve a comprehensive understanding of the challenges and opportunities. We hope that the insights provided in this paper will stimulate an open discussion within the legal NLP community and beyond.

Ethics Statement

It is important to acknowledge that utilizing historical data to train data-driven models may inadvertently introduce biases into the system. For example, Chalkidis et al. (2022b) investigated disparities in classification performance based on factors such as gender, age, and respondent state in human rights litigation. Similar efforts to scrutinize for fairness and bias have been undertaken by Wang et al. (2021); Santosh et al. (2024a); Li et al. (2022). Moreover, recent pre-trained models can inherit biases encoded within their pre-training data. Therefore, any data-driven legal NLP system intended for practical deployment must undergo rigorous scrutiny to ensure compliance with applicable equal treatment and transparency imperatives. This should encompass their performance, behavior, and intended application.

We reiterate the pioneering work in AI & Law by Buchanan and Headrick 1970, which suggested that the computer modeling of legal reasoning would be a fruitful area for research, so as to foster better understanding of legal reasoning and legal argument formation. While we do not advocate for the direct application of predictive systems within courts, the contributions of this paper are intended to facilitate research in this area to enhance transparency, accountability, and explainability. Our goal is to align NLP systems supporting legal practitioners as closely as possible with legal expertise, and to contribute to the discussion around their ethical use.

References

Michael Aikenhead. 1996. Uses and abuses of neural networks in law, the. *Santa Clara Computer & High Tech. LJ*, 12:31.

Latifa Al-Abdulkarim, Katie Atkinson, and Trevor Bench-Capon. 2016a. Accommodating change. *Artificial Intelligence and Law*, 24:409–427.

Latifa Al-Abdulkarim, Katie Atkinson, and Trevor Bench-Capon. 2016b. A methodology for designing systems to reason with legal cases using abstract dialectical frameworks. *Artificial Intelligence and Law*, 24:1–49.

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ computer science*, 2:e93.

Vincent AWMM Aleven. 1997. *Teaching case-based argumentation through a model and examples*. Citeseer.

Basit Ali, Sachin Pawar, Girish Palshikar, Anindita Sinha Banerjee, and Dhirendra Singh. 2023. Legal argument extraction from court judgements using integer linear programming. In *Proceedings of the 10th Workshop on Argument Mining*, pages 52–63.

Basit Ali, Sachin Pawar, Girish Palshikar, and Rituraj Singh. 2022. Constructing a dataset of support and attack relations in legal arguments in court judgements using linguistic rules. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 491–500.

Michal Araszkiewicz. 2021. Critical questions to argumentation schemes in statutory interpretation. *FLAP*, 8(1):291–320.

Kevin D Ashley. 1991. Reasoning with cases and hypotheticals in hypo. *International journal of man-machine studies*, 34(6):753–796.

Kevin D Ashley. 2022. Prospects for legal analytics: some approaches to extracting more meaning from legal texts. *University of Cincinnati Law Review*, 90(4):5.

Kevin D Ashley and Stefanie Brüninghaus. 2009. Automatically classifying case texts and predicting outcomes. *Artificial Intelligence and Law*, 17:125–165.

Katie Atkinson, Trevor Bench-Capon, and Peter McBurney. 2006. Parmenides: facilitating deliberation in democracies. *Artificial Intelligence and Law*, 14:261–275.

Katie Atkinson, Trevor Bench-Capon, Tom Routen, Alejandro Sánchez, Stuart Whittle, Rob Williams, Catriona Wolfenden, and LLP Weightmans. 2019. Implementing angelic designs using logiak. Technical report, Technical Report ULCS-19-002, University of Liverpool.

Trevor Bench-Capon. 1993. Neural networks and open texture. In *Proceedings of the 4th international conference on Artificial intelligence and law*, pages 292–297.

Trevor Bench-Capon and Giovanni Sartor. 2001. Theory based explanation of case law domains: 38. In *Proceedings of the 8th international conference on artificial intelligence and law*, pages 12–21.

Trevor Bench-Capon and Giovanni Sartor. 2003. A model of legal reasoning with cases incorporating theories and values. *Artificial Intelligence*, 150(1-2):97–143.

Trevor JM Bench-Capon and Giovanni Sartor. 2000. Using values and theories to resolve disagreement in law. *Legal knowledge and information systems: Jurix*, pages 73–84.

Donald H Berman and Carole D Hafner. 1991. Incorporating procedural context into a model of case-based legal reasoning. In *Proceedings of the 3rd international conference on Artificial intelligence and law*, pages 12–20.

Donald H Berman and Carole D Hafner. 1993. Representing teleological structure in case-based legal reasoning: the missing link. In *Proceedings of the 4th international conference on Artificial intelligence and law*, pages 50–59.

Donald H Berman and Carole D Hafner. 1995. Understanding precedents in a temporal context of evolving legal doctrine. In *Proceedings of the 5th international conference on Artificial intelligence and law*, pages 42–51.

Vithor Gomes Ferreira Bertalan and Evandro Eduardo Seron Ruiz. 2020. Predicting judicial outcomes in the brazilian legal system using textual features. In *DHandNLP@ PROPOR*, pages 22–32.

Floris Bex. 2011. *Arguments, stories and criminal evidence: A formal hybrid theory*, volume 92. Springer Science & Business Media.

Floris Bex, Henry Prakken, Chris Reed, and Douglas Walton. 2003. Towards a formal account of reasoning about evidence: argumentation schemes and generalisations. *Artificial Intelligence and Law*, 11:125–165.

Laurent Bochereau, Danièle Bourcier, and Paul Bourgine. 1991. Extracting legal knowledge by means of a multilayer neural network application to municipal jurisprudence. In *Proceedings of the 3rd international conference on Artificial intelligence and law*, pages 288–296.

L Karl Branting. 1991. Building explanations from rules and structured cases. *International journal of man-machine studies*, 34(6):797–837.

L Karl Branting. 1993. A computational model of ratio decidendi. *Artificial intelligence and law*, 2:1–31.

L Karl Branting, Craig Pfeifer, Bradford Brown, Lisa Ferro, John Aberdeen, Brandy Weiss, Mark Pfaff, and Bill Liao. 2021. Scalable and explainable legal prediction. *Artificial Intelligence and Law*, 29:213–238.

Gerhard Brewka, Stefan Ellmauthaler, Hannes Strass, Johannes Peter Wallner, and Stefan Woltran. 2013. Abstract dialectical frameworks revisited. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 803–809.

Stefanie Bruninghaus and Kevin D Ashley. 2003. Predicting outcomes of case based legal arguments. In *Proceedings of the 9th international conference on Artificial intelligence and law*, pages 233–242.

Bruce G Buchanan and Thomas E Headrick. 1970. Some speculation about artificial intelligence and legal reasoning. *Stan. L. Rev.*, 23:40.

Martin Caminada and Leila Amgoud. 2007. On the evaluation of argumentation formalisms. *Artificial Intelligence*, 171(5-6):286–310.

Juan Bayón Carlos. 2001. Why is legal reasoning defeasible? In *Pluralism and law*, pages 327–346. Springer.

Ilias Chalkidis. 2023. Chatgpt may pass the bar exam soon, but has a long way to go for the lexglue benchmark. *arXiv preprint arXiv:2304.12202*.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in english. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. Paragraph-level rationale extraction through regularization: A case study on european court of human rights cases. *arXiv preprint arXiv:2103.13084*.

Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Martin Katz, and Anders Søgaard. 2023. Lexfiles and legallama: Facilitating english multinational legal language model development. *arXiv preprint arXiv:2305.07507*.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022a. 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.

Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Felix Schwemer, and Anders Søgaard. 2022b. Fairlex: A multilingual benchmark for evaluating fairness in legal text processing. *arXiv preprint arXiv:2203.07228*.

Alison Chorley and Trevor Bench-Capon. 2004. Support for constructing theories in case law domains. In *International Conference on Database and Expert Systems Applications*, pages 508–517. Springer.

Alison Chorley and Trevor Bench-Capon. 2005a. Agatha: Using heuristic search to automate the construction of case law theories. *Artificial Intelligence and Law*, 13:9–51.

Alison Chorley and Trevor Bench-Capon. 2005b. An empirical investigation of reasoning with legal cases through theory construction and application. *Artificial Intelligence and Law*, 13:323–371.

Wentao Deng, Jiahuan Pei, Keyi Kong, Zhe Chen, Furu Wei, Yujun Li, Zhaochun Ren, Zhumin Chen, and Pengjie Ren. 2023. Syllogistic reasoning for legal judgment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13997–14009.

Qian Dong and Shuzi Niu. 2021. Legal judgment prediction via relational learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 983–992.

Stella Douka, Hadi Abdine, Michalis Vazirgiannis, Rajaa El Hamdani, and David Restrepo Amariles. 2021. Juribert: A masked-language model adaptation for french legal text. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 95–101.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357.

Eisner v. Macomber. 252 U.S. 189, 207, 40 S.Ct. 189, 64 L.Ed. 521 (1920).

Mohamed Elaraby, Huihui Xu, Morgan Gray, Kevin D Ashley, and Diane Litman. 2024. Adding argumentation into human evaluation of long document abstractive summarization: A case study on legal opinions. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval)@ LREC-COLING 2024*, pages 28–35.

Yi Feng, Chuanyi Li, and Vincent Ng. 2022. Legal judgment prediction via event extraction with constraints. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 648–664.

Pedro Miguel Freitas and Luís Mendes Gomes. 2023. Does chatgpt pass the brazilian bar exam? In *EPIA Conference on Artificial Intelligence*, pages 131–141. Springer.

Leilei Gan, Kun Kuang, Yi Yang, and Fei Wu. 2021. Judgment prediction via injecting legal knowledge into neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12866–12874.

Leilei Gan, Baokui Li, Kun Kuang, Yi Yang, and Fei Wu. 2022. Exploiting contrastive learning and numerical evidence for improving confusing legal judgment prediction. *arXiv preprint arXiv:2211.08238*.

Anne von der Lieth Gardner. 1987. *An artificial intelligence approach to legal reasoning*. MIT press.

Thomas F Gordon. 1993. The pleadings game: An exercise in computational dialectics. *Artificial Intelligence and Law*, 2:239–292.

Thomas F Gordon and Douglas Walton. 2009. Legal reasoning with argumentation schemes. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 137–146.

Matthias Grabmair. 2016. *Modeling purposive legal argumentation and case outcome prediction using argument schemes in the value judgment formalism*. Ph.D. thesis, University of Pittsburgh.

Matthias Grabmair. 2017. Predicting trade secret case outcomes using argument schemes and learned quantitative value effect tradeoffs. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pages 89–98.

Matthias Grabmair and Kevin D Ashley. 2011. Facilitating case comparison using value judgments and intermediate legal concepts. In *Proceedings of the 13th international conference on Artificial intelligence and law*, pages 161–170.

Matthias Grabmair, Kevin D Ashley, Ran Chen, Preethi Sureshkumar, Chen Wang, Eric Nyberg, and Vern R Walker. 2015. Introducing luima: an experiment in legal conceptual retrieval of vaccine injury decisions using a uima type system and tools. In *Proceedings of the 15th international conference on artificial intelligence and law*, pages 69–78.

Morgan Gray, Jaromir Savelka, Wesley Oliver, and Kevin Ashley. 2023. Automatic identification and empirical analysis of legally relevant factors. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 101–110.

Morgan A Gray, Jaromir Savelka, Wesley M Oliver, and Kevin D Ashley. 2024. 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.

Katie Greenwood, Trevor Bench Capon, and Peter McBurney. 2003. Towards a computational account of persuasion in law. In *Proceedings of the 9th international conference on artificial intelligence and law*, pages 22–31.

Giulia Grundler, Piera Santin, Andrea Galassi, Federico Galli, Francesco Godano, Francesca Lagioia, Elena Palmieri, Federico Ruggeri, Giovanni Sartor, and Paolo Torroni. 2022. Detecting arguments in cjeu decisions on fiscal state aid. In *Proceedings of the 9th Workshop on Argument Mining*, pages 143–157.

Neel Guha, Julian Nyarko, Daniel E Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N Rockmore, et al. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *arXiv preprint arXiv:2308.11462*.

Ivan Habernal, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Indra Spiecker genannt Döhmann, and Christoph Burchard. 2023. Mining legal arguments in court decisions. *Artificial Intelligence and Law*, pages 1–38.

Jaap C Hage, Ronald Leenes, and Arno R Lodder. 1993. Hard cases: a procedural approach. *Artificial intelligence and law*, 2:113–167.

Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. 2021. Do language models have beliefs? methods for detecting, updating, and visualizing model beliefs. *arXiv preprint arXiv:2111.13654*.

John Henderson and Trevor Bench-Capon. 2019. Describing the development of case law. In *Proceedings of the seventeenth international conference on artificial intelligence and law*, pages 32–41.

Nils Holzenberger and Benjamin Van Durme. 2021. Factoring statutory reasoning as language understanding challenges. In *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)*, pages 2742–2758.

Nils Holzenberger and Benjamin Van Durme. 2023. Connecting symbolic statutory reasoning with legal information extraction. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 113–131.

Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. Few-shot charge prediction with discriminative legal attributes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 487–498.

Yunyun Huang, Xiaoyu Shen, Chuanyi Li, Jidong Ge, and Bin Luo. 2021. Dependency learning for legal judgment prediction with a unified text-to-text transformer. *arXiv preprint arXiv:2112.06370*.

Dan Hunter. 1994. Looking for law in all the wrong places: Legal theory and legal neural networks. *Legal knowledge based systems: The relation with legal theory*, pages 55–64.

Wonseok Hwang, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo. 2022. A multi-task benchmark for korean legal language understanding and judgement prediction. *Advances in Neural Information Processing Systems*, 35:32537–32551.

Cong Jiang and Xiaolei Yang. 2023. Legal syllogism prompting: Teaching large language models for legal judgment prediction. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 417–421.

Xiaoxi Kang, Lizhen Qu, Lay-Ki Soon, Adnan Trakic, Terry Yue Zhuo, Patrick Charles Emerton, and Genevieve Grant. 2023. Can chatgpt perform reasoning using the irac method in analyzing legal scenarios like a lawyer? In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Daniel Martin Katz, Michael J Bommarito, and Josh Blackman. 2017. A general approach for predicting the behavior of the supreme court of the united states. *PloS one*, 12(4):e0174698.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2023a. Gpt-4 passes the bar exam. *Available at SSRN 4389233*.

Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael J Bommarito II. 2023b. Natural language processing in the legal domain. *arXiv preprint arXiv:2302.12039*.

Aaron Russell Kaufman, Peter Kraft, and Maya Sen. 2019. Improving supreme court forecasting using boosted decision trees. *Political Analysis*, 27(3):381–387.

Vladimir Khairoulline. 2007. The discourse of court interpreting: Discourse practices of the law, the witness and the interpreter.

Fred Kort. 1957. Predicting supreme court decisions mathematically: A quantitative analysis of the "right to counsel" cases. *American Political Science Review*, 51(1):1–12.

André Lage-Freitas, Héctor Allende-Cid, Orivaldo Santana, and Lívia Oliveira-Lage. 2022. Predicting brazilian court decisions. *PeerJ Computer Science*, 8:e904.

Yanjun Li, Huan Huang, Qiang Geng, Xinwei Guo, and Yuyu Yuan. 2022. Fairness measures of machine learning models in judicial penalty prediction. *Journal of Internet Technology*, 23(5):1109–1116.

Dugang Liu, Weihao Du, Lei Li, Weike Pan, and Zhong Ming. 2022. Augmenting legal judgment prediction with contrastive case relations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2658–2667.

Yifei Liu, Yiquan Wu, Yating Zhang, Changlong Sun, Weiming Lu, Fei Wu, and Kun Kuang. 2023. Ml-ljp: Multi-law aware legal judgment prediction. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1023–1034.

Zhenyu Liu and Huanhuan Chen. 2017. A predictive performance comparison of machine learning models for judicial cases. In *2017 IEEE Symposium series on computational intelligence (SSCI)*, pages 1–6. IEEE.

Ronald Prescott Loui and Jeff Norman. 1995. Rationales and argument moves. *Artificial Intelligence and Law*, 3:159–189.

Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. Learning to predict charges for criminal cases with legal basis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2727–2736.

Luyao Ma, Yating Zhang, Tianyi Wang, Xiaozhong Liu, Wei Ye, Changlong Sun, and Shikun Zhang. 2021. Legal judgment prediction with multi-stage case representation learning in the real court setting. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 993–1002.

Fabrizio Macagno, Douglas Walton, and Chris Reed. 2017. Argumentation schemes. history, classifications, and computational applications. *History, Classifications, and Computational Applications (December 23, 2017). Macagno, F., Walton, D. & Reed, C*, pages 2493–2556.

Ejan Mackaay and Pierre Robillard. 1974. *Predicting judicial decisions: The nearest neighbour rule and visual representation of case patterns*.

Robert Mahari, Dominik Stammbach, Elliott Ash, and Alex'Sandy' Pentland. 2023. The law and nlp: Bridging disciplinary disconnects. *arXiv preprint arXiv:2310.14346*.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. Ildc for cjpe: Indian legal documents corpus for court judgment prediction and explanation. In *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)*, pages 4046–4062.

Juliano Maranhao, Edelcio G de Souza, and Giovanni Sartor. 2021. A dynamic model for balancing values. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 89–98.

Mihai Masala, Radu Cristian Alexandru Iacob, Ana Sabina Uban, Marina Cidota, Horia Velicu, Traian Rebedea, and Marius Popescu. 2021. jurbert: A romanian bert model for legal judgement prediction. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 86–94.

L Thorne McCarty. 1976. Reflections on taxman: An experiment in artificial intelligence and legal reasoning. *Harvard Law Review*, 90:837.

L Thorne McCarty. 1995. An implementation of eisner v. macomber. In *Proceedings of the 5th international conference on Artificial intelligence and law*, pages 276–286.

Masha Medvedeva and Pauline Mcbride. 2023. Legal judgment prediction: If you are going to do it, do it right. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 73–84.

Masha Medvedeva, Ahmet Üstün, Xiao Xu, Michel Vols, and Martijn Wieling. 2021. Automatic judgement forecasting for pending applications of the european court of human rights. In *ASAIL/LegalAIIA@ ICAIL*, pages 12–23.

Masha Medvedeva, Michel Vols, and Martijn Wieling. 2020. Using machine learning to predict decisions of the european court of human rights. *Artificial Intelligence and Law*, 28(2):237–266.

Masha Medvedeva, Martijn Wieling, and Michel Vols. 2023. Rethinking the field of automatic prediction of court decisions. *Artificial Intelligence and Law*, 31(1):195–212.

Sanjay Modgil. 2009. Reasoning about preferences in argumentation frameworks. *Artificial intelligence*, 173(9-10):901–934.

Ankan Mullick, Abhilash Nandy, Manav Nitin Kapadnis, Sohan Patnaik, R Raghav, and Roshni Kar. 2022. An evaluation framework for legal document summarization. *arXiv preprint arXiv:2205.08478*.

Jack Mumford, Katie Atkinson, and Trevor Bench-Capon. 2022. Reasoning with legal cases: A hybrid adf-ml approach. In *Legal Knowledge and Information Systems*, pages 93–102. IOS Press.

Jack Mumford, Katie Atkinson, and Trevor Bench-Capon. 2023a. Combining a legal knowledge model with machine learning for reasoning with legal cases. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 167–176.

Jack Mumford, Katie Atkinson, and Trevor Bench-Capon. 2023b. Human performance on the ai legal case verdict classification task. *Frontiers in Artificial Intelligence and Applications*, 379:359–364.

Robert Muthuri, Guido Boella, Joris Hulstijn, Sara Capecchi, and Llio Humphreys. 2017. Compliance patterns: harnessing value modeling and legal interpretation to manage regulatory conversations. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pages 139–148.

Stuart S Nagel. 1963. Applying correlation analysis to case prediction. *Tex. L. Rev.*, 42:1006.

Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. 2022. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35.

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel E Ho. 2023. Multilegalpile: A 689gb multilingual legal corpus. *arXiv preprint arXiv:2306.02069*.

Mark K Osbeck and Michael Gilliland. 2018. Outcome prediction in the practice of law. *Foresight: The International Journal of Applied Forecasting*, (50):42–48.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107.

Anandeep S Pannu. 1995. Using genetic algorithms to inductively reason with cases in the legal domain. In *Proceedings of the 5th international conference on Artificial intelligence and law*, pages 175–184.

Chaim Perelman and Lucie Olbrechts-Tyteca. 1969. The new rhetoric: a treatise on argumentation, trans. *John Wilkinson and Purcell Weaver (Notre Dame, IN: University of Notre Dame Press, 1969)*, 19.

Lothar Philipps. 1989. A neural network to identify legal precedents. In *Proceedings of the 9th Symposium on Legal Data Processing in Europe. Council of Europes*, pages 99–106.

Prakash Poudyal, Jaromír Šavelka, Aagje Ieven, Marie Francine Moens, Teresa Goncalves, and Paulo Quaresma. 2020. Echr: Legal corpus for argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, pages 67–75.

Henry Prakken. 1995. From logic to dialectics in legal argument. In *Proceedings of the 5th international conference on artificial intelligence and law*, pages 165–174.

Henry Prakken. 2010. An abstract framework for argumentation with structured arguments. *Argument & Computation*, 1(2):93–124.

Henry Prakken. 2012. Reconstructing popov v. hayashi in a framework for argumentation with structured arguments and dungean semantics. *Artificial Intelligence and law*, 20:57–82.

Henry Prakken and Giovanni Sartor. 1997. A dialectical model of assessing conflicting arguments in legal reasoning. *Logical models of legal argumentation*, pages 175–211.

Henry Prakken and Giovanni Sartor. 1998. Modelling reasoning with precedents in a formal dialogue game. *Judicial applications of artificial intelligence*, pages 127–183.

Henry Prakken, Adam Wyner, Trevor Bench-Capon, and Katie Atkinson. 2015. A formalization of argumentation schemes for legal case-based reasoning in aspic+. *Journal of Logic and Computation*, 25(5):1141–1166.

Edwina L Rissland, Kevin D Ashley, and Ronald Prescott Loui. 2003. Ai and law: A fruitful synergy. *Artificial Intelligence*, 150(1-2):1–15.

Edwina L Rissland and David B Skalak. 1991. Cabaret: rule interpretation in a hybrid architecture. *International journal of man-machine studies*, 34(6):839–887.

Edwina L Rissland, David B Skalak, and M Timur Friedman. 1996. Bankxx: Supporting legal arguments through heuristic retrieval. *Artificial Intelligence and Law*, 4(1):1–71.

Edwina L Rissland, David B Skalak, and M Timur Friedman. 1997. Evaluating a legal argument program: The bankxx experiments. *Artificial Intelligence and Law*, 5:1–74.

Edwina L Rissland and Xiaoxi Xu. 2011. Catching gray cygnets: an initial exploration. In *Proceedings of the 13th international conference on artificial intelligence and law*, pages 151–160.

Theodore W Ruger, Pauline T Kim, Andrew D Martin, and Kevin M Quinn. 2004. The supreme court forecasting project: Legal and political science approaches to predicting supreme court decisionmaking. *Colum. L. Rev.*, 104:1150.

Olivier Salaün, Aurore Troussel, Sylvain Longhais, Hannes Westermann, Philippe Langlais, and Karim Benyekhlef. 2022. Conditional abstractive summarization of court decisions for laymen and insights from human evaluation. In *Legal Knowledge and Information Systems*, pages 123–132. IOS Press.

Santosh, Nina T.Y.S.S, Baumgartner, Matthias Stürmer, Matthias Grabmair, Joel Niklaus, et al. 2024a. Towards explainability and fairness in swiss judgement predicti on: Benchmarking on a multilingual dataset. *arXiv preprint arXiv:2402.17013*.

TYS Santosh, Marcel Perez San Blas, Phillip Kemper, and Matthias Grabmair. 2023a. Leveraging task dependency and contrastive learning for case outcome classification on european court of human rights cases. *arXiv preprint arXiv:2302.00768*.

TYS Santosh, Oana Ichim, and Matthias Grabmair. 2023b. Zero-shot transfer of article-aware legal outcome classification for european court of human rights cases. *arXiv preprint arXiv:2302.00609*.

TYS Santosh, Tuan-Quang Vuong, and Matthias Grabmair. 2024b. Chronoslex: Time-aware incremental training for temporal generalization of legal classification tasks. *arXiv preprint arXiv:2405.14211*.

TYSS Santosh, Mohamed Hesham Elganayni, Stanisław Sójka, and Matthias Grabmair. 2024c. Incorporating precedents for legal judgement prediction on european court of human rights cases. *arXiv preprint arXiv:2409.18644*.

Tyss Santosh, Shanshan Xu, Oana Ichim, and Matthias Grabmair. 2022. Deconfounding legal judgment prediction for european court of human rights cases towards better alignment with experts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1120–1138.

Jaromir Savelka, Kevin D Ashley, Morgan A Gray, Hannes Westermann, and Huihui Xu. 2023. Can gpt-4 support analysis of textual data in tasks requiring highly specialized domain expertise? *arXiv preprint arXiv:2306.13906*.

JURI SAYS. 2020. Prediction system for the european court of human rights. In *Legal Knowledge and Information Systems: JURIX 2020: The Thirty-third Annual Conference, Brno, Czech Republic, December 9-11, 2020*, volume 334, page 277. IOS Press.

Jeffrey A Segal. 1984. Predicting supreme court cases probabilistically: The search and seizure cases, 1962-1981. *American Political Science Review*, 78(4):891–900.

Marek J. Sergot, Fariba Sadri, Robert A. Kowalski, Frank Kriwaczek, Peter Hammond, and H Terese Cory. 1986. The british nationality act as a logic program. *Communications of the ACM*, 29(5):370–386.

Mehmet Fatih Sert, Engin Yıldırım, and İrfan Haşlak. 2021. Using artificial intelligence to predict decisions of the turkish constitutional court. *Social Science Computer Review*, page 08944393211010398.

Rafe Athar Shaikh, Tirath Prasad Sahu, and Veena Anand. 2020. Predicting outcomes of legal cases based on legal factors using classifiers. *Procedia Computer Science*, 167:2393–2402.

Ruihao Shui, Yixin Cao, Xiang Wang, and Tat-Seng Chua. 2023. A comprehensive evaluation of large language models on legal judgment prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7337–7348.

Andrew Stranieri, John Zeleznikow, Mark Gawler, and Bryn Lewis. 1999. A hybrid rule–neural approach for the automation of legal reasoning in the discretionary domain of family law in australia. *Artificial intelligence and law*, 7(2-3):153–183.

Benjamin Strickson and Beatriz De La Iglesia. 2020. Legal judgement prediction for uk courts. In *Proceedings of the 3rd International Conference on Information Science and Systems*, pages 204–209.

Octavia-Maria Şulea, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P Dinu, and Josef van Genabith. 2017a. Exploring the use of text classification in the legal domain.

Octavia-Maria Şulea, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. 2017b. Predicting the law area and decisions of french supreme court cases. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 716–722.

Elizabeth C Tippett, Charlotte S Alexander, Karl Branting, Paul Morawski, Carlos Balhana, Craig Pfeifer, and Sam Bayer. 2021. Does lawyering matter? predicting judicial decisions from legal briefs, and what that means for access to justice. *Tex. L. Rev.*, 100:1157.

Stephen Edelston Toulmin. 1958. The uses of argument.

Dietrich Trautmann. 2023. Large language model prompt chaining for long legal document classification. *arXiv preprint arXiv:2308.04138*.

Dietrich Trautmann, Alina Petrova, and Frank Schilder. 2022. Legal prompt engineering for multilingual legal judgement prediction. *arXiv preprint arXiv:2212.02199*.

Santosh Tyss, Oana Ichim, and Matthias Grabmair. 2023a. Zero-shot transfer of article-aware legal outcome classification for european court of human rights cases. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 593–605.

Santosh Tyss, Marcel Perez San Blas, Phillip Kemper, and Matthias Grabmair. 2023b. Leveraging task dependency and contrastive learning for case outcome classification on european court of human rights cases. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1103–1103.

Sebastian Urbina. 2002. *Legal method and the rule of law*, volume 59. Springer Science & Business Media.

Josef Valvoda, Ryan Cotterell, and Simone Teufel. 2023. On the role of negative precedent in legal outcome prediction. *Transactions of the Association for Computational Linguistics*, 11:34–48.

Shaurya Vats, Atharva Zope, Somsubhra De, Anurag Sharma, Upal Bhattacharya, Shubham Nigam, Shouvik Guha, Koustav Rudra, and Kripabandhu Ghosh. 2023. Llms–the good, the bad or the indispensable?: A use case on legal statute prediction and legal judgment prediction on indian court cases. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12451–12474.

Bart Verheij. 2001. Legal decision making as dialectical theory construction with argumentation schemes. In *proceedings of the 8th International Conference on Artificial Intelligence and Law*, pages 225–226.

Bart Verheij. 2016. Formalizing value-guided argumentation for ethical systems design. *Artificial Intelligence and Law*, 24:387–407.

Michael Benedict L Virtucio, Jeffrey A Aborot, John Kevin C Abonita, Roxanne S Avinante, Rother Jay B Copino, Michelle P Neverida, Vanesa O Osiana, Elmer C Peramo, Joanna G Syjuco, and Glenn Brian A Tan. 2018. Predicting decisions of the philippine supreme court using natural language processing and machine learning. In *2018 IEEE 42nd annual computer software and applications conference (COMPSAC)*, volume 2, pages 130–135. IEEE.

Bernhard Waltl, Georg Bonczek, Elena Scepankova, Jörg Landthaler, and Florian Matthes. 2017. Predicting the outcome of appeal decisions in germany's tax law. In *International conference on electronic participation*, pages 89–99. Springer.

Douglas N Walton. 1996. Argumentation schemes for presumptive reasoning.

Yuzhong Wang, Chaojun Xiao, Shirong Ma, Haoxi Zhong, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2021. Equality before the law: Legal judgment consistency analysis for fairness. *arXiv preprint arXiv:2103.13868*.

Donald A Waterman and Mark Peterson. 1980. Rule-based models of legal expertise. In *AAAI*, volume 1, pages 272–275.

Adam Wyner and Trevor Bench-Capon. 2009. Modelling judicial context in argumentation frameworks. *Journal of Logic and Computation*, 19(6):941–968.

Adam Wyner, Raquel Mochales-Palau, Marie-Francine Moens, and David Milward. 2010. *Approaches to text mining arguments from legal cases*. Springer.

Adam Zachary Wyner, Trevor J. M. Bench-Capon, and Katie Atkinson. 2007. Arguments, values and baseballs: Representation of popov v. hayashi. In *Legal Knowledge and Information Systems - JURIX 2007: The Twentieth Annual Conference on Legal Knowledge and Information Systems, Leiden, The Netherlands, 12-15 December 2007*, volume 165 of *Frontiers in Artificial Intelligence and Applications*, pages 151–160. IOS Press.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84.

Huihui Xu and Kevin D. Ashley. 2022. Multi-granularity argument mining in legal texts. In *International Conference on Legal Knowledge and Information Systems*.

Shanshan Xu, Oana Ichim, Isabella Risini, Barbara Plank, Matthias Grabmair, et al. 2023. From dissonance to insights: Dissecting disagreements in rationale construction for case outcome classification. *arXiv preprint arXiv:2310.11878*.

Shanshan Xu, TYS Santosh, Oana Ichim, Barbara Plank, and Matthias Grabmair. 2024. Through the lens of split vote: Exploring disagreement, difficulty and calibration in legal case outcome classification. *arXiv preprint arXiv:2402.07214*.

Wenmian Yang, Weijia Jia, Xiaojie Zhou, and Yutao Luo. 2019. Legal judgment prediction via multi-perspective bi-feedback network. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 4085–4091.

Fangyi Yu, Lee Quartey, and Frank Schilder. 2022. Legal prompting: Teaching a language model to think like a lawyer. *arXiv preprint arXiv:2212.01326*.

Fangyi Yu, Lee Quartey, and Frank Schilder. 2023. Exploring the effectiveness of prompt engineering for legal reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13582–13596.

Linan Yue, Qi Liu, Binbin Jin, Han Wu, Kai Zhang, Yanqing An, Mingyue Cheng, Biao Yin, and Dayong Wu. 2021. Neurjudge: A circumstance-aware neural framework for legal judgment prediction. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 973–982.

Han Zhang, Zhicheng Dou, Yutao Zhu, and Ji-Rong Wen. 2023. Contrastive learning for legal judgment prediction. *ACM Transactions on Information Systems*, 41(4):1–25.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. 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.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3540–3549.

Haoxi Zhong, Yuzhong Wang, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020a. Iteratively questioning and answering for interpretable legal judgment prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1250–1257.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020b. How does nlp benefit legal system: A summary of legal artificial intelligence. *arXiv preprint arXiv:2004.12158*.