δ-Stance: A Large-Scale Real World Dataset of Stances in Legal Argumentation

Ankita Gupta **Douglas Rice** **Brendan O'Connor**
University of Massachusetts Amherst
{ankitagupta,brenocon}@cs.umass.edu, drrice@umass.edu

Abstract

We present δ-Stance, a large-scale dataset of stances involved in legal argumentation. δ-Stance contains stance-annotated argument pairs, semi-automatically mined from millions of examples of U.S. judges citing precedent in context using citation signals. The dataset aims to facilitate work on the *legal argument stance classification task*, which involves assessing whether a *case summary* strengthens or weakens a legal *argument* (polarity) and to what extent (intensity). To assess the complexity of this task, we evaluate various existing NLP methods, including zero-shot prompting proprietary large language models (LLMs), and supervised fine-tuning of smaller open-weight language models (LMs) on δ-Stance. Our findings reveal that although prompting proprietary LLMs can help predict stance polarity, supervised model fine-tuning on δ-Stance is necessary to distinguish intensity. We further find that alternative strategies such as domain-specific pretraining and zero-shot prompting using masked LMs remain insufficient. Beyond our dataset's utility for the legal domain, we further find that fine-tuning small LMs on δ-Stance improves their performance in other domains. Finally, we study how temporal changes in signal definition can impact model performance, highlighting the importance of careful data curation for downstream tasks by considering the historical and sociocultural context. We publish the associated dataset[1] to foster further research on legal argument reasoning.

1 Introduction

Argumentation plays an important role in many areas (e.g., science, law, governance, and journalism), where decision-making requires professionals to sort through contradictory evidence to construct well-reasoned, grounded arguments that acknowledge (and sometimes rebut) alternative viewpoints (Palau and Moens, 2009). However, manual interpretation of such arguments and associated evidence is often challenging, where practitioners must navigate vast amounts of information. For example, U.S. federal courts have produced approximately 1.7 million published judicial opinions, resulting in millions of potentially relevant precedents that could be cited in new cases (Mahari et al., 2024).

Current tools offer limited support for this work. LLM-based chatbots or writing assistants can generate generic text but lack strong evidence grounding, resulting in fabricated facts and citations (Weiser, 2023). Contemporary search engines and retrieval methods can suggest *which* documents are relevant, but analysts must still manually determine *how* each document connects to their argument. To help automate the drafting and analysis of legal arguments at scale, we propose the task of *legal argument stance classification* that aims to assess whether a *case summary* strengthens or weakens a legal *argument* (polarity) and to what extent (intensity).

Our task can be useful for many applications. Identifying precedents relevant to an author's argument is fundamental to legal practice. While previous methods help retrieve all relevant precedents, the results can be superficial (see, e.g., Chapman, 2024) and require significant manual engagement. In contrast, a legal argument stance classifier can provide deeper analysis by helping to infer argument stances–whether the retrieved precedents strengthen or weaken the argument. The classifier can also be integrated as a tool with function-calling LLMs (Schick et al., 2023) to identify a retrieved precedent's relationship to a legal argument or evaluate LLM-generated responses by examining the relevance of cited cases to the overall argument.

Further, such technology could be incorporated into writing assistance systems to help authors better articulate the connection between their work and prior research (Luu et al., 2021). For instance, it can help assess an argument's strength by visualizing

[1] https://github.com/slanglab/deltastance

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 31450–31467

stance distribution among retrieved precedents (e.g., if most are strengtheners, the context may be a well-supported argument), identify flaws in an argument by retrieving only weakeners, or detect mis-citations in legal texts. Our task can also help identify relationships between cases that do not cite each other (e.g., contemporaneous cases intentionally or unintentionally omitted as precedents).

In this work, we utilize citations to precedents found in judicial opinions and legal writing conventions to semi-automatically mine a dataset to support our task. Specifically, we curate and release δ-Stance, a large-scale dataset with millions of annotated argument pairs, each structured as a triple containing an *argument context*, a *case summary*, and a *stance value*. These triples are extracted from U.S. judicial opinions accessed via the Caselaw Access Project.[2] To identify the triples, we exploit systematic regularities in prescriptive citation signal words used by experts as part of routine legal practice (Landes et al., 1998). Unlike prior methods that rely on collecting costly, time-consuming annotations by experts (Habernal et al., 2024; Poudyal et al., 2020) or using educational material (Bongard et al., 2022), our dataset uses a semi-automated process to mine naturally occurring expert-annotated triples grounded in real-world historical legal practice.

Our major contributions include:

- We detail the construction of δ-Stance (§3) and provide dataset statistics and underlying socio-cultural context (§4).
- We evaluate various stance classification approaches (§5), including zero-shot prompting and supervised fine-tuning on our task. Our results show that prompting proprietary LLMs can effectively identify stance polarity, though supervised fine-tuning is necessary to distinguish the intensity, underscoring the complexity of legal stance classification (§5.2).
- We further conduct additional analyses including alternative zero-shot approaches using masked language models (§5.3), and examine the impact of domain pre-training (§5.4) and temporal changes in signal definitions (§5.5).
- Finally, we show δ-Stance's broader applicability beyond the legal domain by demonstrating that LMs trained on this dataset improve their argument reasoning abilities in other domains (§5.6).

Overall, our work demonstrates an example of using normative standards to harvest useful linguistic data at scale, enabling the development of automated stance/reasoning models and conducting empirical legal studies. While δ-Stance is based on the U.S. legal citation system, our data curation methodology could be adapted to any other jurisdiction employing citation signals in their legal writing (e.g., Australia,[3] Canada[4]), offering promising avenues for future research.

2 Related Work

Our work engages with the following research areas:

NLP for law. Legal text presents unique NLP challenges, including extraordinarily long documents, technical language requiring specific expertise, intentional ambiguity (Chalkidis et al., 2022; Li et al., 2023; Dai et al., 2022), and complex argument structures (Habernal et al., 2024). Moreover, legal texts are also quite diverse in subject matter—from local property codes to presidential powers during national emergencies (Katz et al., 2023).

Legal reasoning. Several benchmarks evaluate legal reasoning in LLMs (Guha et al., 2024; Chalkidis et al., 2022), including statutory reasoning (Holzenberger et al., 2020), identifying the best supporting statement for an argument (Zheng et al., 2021; Liang et al., 2023) or the most likely continuation of a given argument (Chlapanis et al., 2024). Guha et al. (2024) introduced a comprehensive benchmark following the IRAC framework, with similar efforts emerging for other jurisdictions (Joshi et al., 2024; Niklaus et al., 2023; Fei et al., 2024). Our work complements these efforts by focusing on the fine-grained stances among legal texts as expressed through citation signals, an essential aspect of legal reasoning that has not been addressed in prior benchmarks.

Legal retrieval. Prior work has developed datasets and methods for the retrieval of supporting cases (Goebel et al., 2023) or paragraphs (Mahari et al., 2024; Goebel et al., 2024) for given queries. Our work augments this by predicting stances for retrieved cases or assisting in retrieving cases of specific stances (e.g., weakeners).

Other semantic tasks. Our work relates closely to *case-based reasoning* (CBR) (Kolodner, 1992;

[2] https://case.law/

[3] https://law.unimelb.edu.au/mulr/aglc/about

[4] https://lawjournal.mcgill.ca/cite-guide/

Aamodt and Plaza, 1994; Das et al., 2021; Ashley and Rissland, 1987), a computational framework for retrieving and reusing solutions from prior experiences to interpret new situations. The US legal system is a natural domain for studying CBR as it operates under the principle of *stare decisis*, Latin for 'to stand by things decided.' Accordingly, our dataset can be useful for studying CBR, where the citation signals help assess the relevance of retrieved precedents. Our task is also related to *defeasible reasoning*, in which conclusions can be revised or withdrawn as new information becomes available (Koons, 2022; Rudinger et al., 2020). In legal argumentation, conclusions may be strengthened or weakened by different precedents, with citation signals indicating how each precedent affects the argument. Further, the work can be viewed as a practical application of the *ordinal entailment task* (Zhang et al., 2016), where the classes have a natural ordering among them, as do our citations.

Our task can also be viewed as a fine-grained *argument relation classification task* (Lawrence and Reed, 2020), in which case summaries serve as reasons (positive stance) or rebuttals (negative stance) to the claims, useful for identifying the most convincing legal arguments (Habernal and Gurevych, 2016). The work also relates to *stance detection*. In our project, case summaries serve as stance-takers, similar to Mohtarami et al. (2018)), and claims embedded in argument contexts are stance objects (Mohammad et al., 2016). Unlike affective or epistemic relations (Biber and Finegan, 1989), our stance relations are argumentative.

Data collection for sociocultural analysis. Our work is also related to ongoing efforts on thoughtful curation of sociocultural data in machine learning (Jo and Gebru, 2020; Dodge et al., 2021). For instance, we leverage the CAP's systematic digitization of the US case law, which draws from a well-defined scholarly universe of US legal practitioners, exemplifying data curation as archival science (Jo and Gebru, 2020) while preserving sociocultural context. Prior work has also highlighted the challenges in collecting historical corpora (e.g., Google Books corpus) and using them for sociocultural analyses, noting that changes in corpus composition can significantly impact the validity of analyses (Pechenick et al., 2015; Schmidt et al., 2021). Our study in §5.5 on the effect of temporal changes in signal definition on model performance exemplifies this aspect.

3 δ-Stance Dataset

3.1 Corpus Description

Caselaw Access Project (CAP). We analyze legal arguments drawn from 7 million published judicial opinions from U.S. federal and state courts from Harvard's CAP, which provides access to raw opinion texts along with opinion metadata (the court, the decision date, etc.). The CAP collection cases date back to 1658 and include official, book-published U.S. case law through 2020. It encompasses judicial opinions from U.S. federal courts—the U.S. Supreme Court, 13 federal appellate courts, and 94 district courts—and U.S. state courts, including State Supreme Courts, Intermediate Courts of Appeals, and State trial courts.

Legal citation writing conventions. Legal authors must adhere to a rigid set of citation guidelines outlined in the *Bluebook: A Uniform System of Citation*, an authoritative guide for legal citation in the U.S. The *Bluebook* provides a comprehensive set of legal writing rules to ensure consistency and clarity in legal writing and has seen widespread adoption in the legal community.[5] As Gallacher (2006) puts it, *"the Bluebook is sometimes referred to as the "Bible" of legal citation,"* highlighting its significance in the legal system.

We draw on two key features of legal citation writing conventions, as prescribed in the *Bluebook*, to curate δ-`Stance`. First, the authors must explicitly state whether a cited case supports (or weakens) their argument and state the level of support provided by the case via a prefix, commonly known to legal practitioners as *citation signals* or *introductory signals*. The Bluebook also provides a comprehensive list of *citation signals* along with their definitions, outlining the appropriate contexts for their use. Second, authors are also required to explain the relevance of each cited precedent in a brief *parenthetical*.

3.2 Argument Stance Classification Task

Based on the above legal writing conventions, we interpret the following *argument stance classification* task:

[5] In addition to the Bluebook, alternative citation guides have emerged, most notably the *ALWD Citation Manual: A Professional System of Citation*, first released in 2000. Despite alternatives, the Bluebook continues to dominate citation instruction (Ryan, 2022). Critically for our work, the citation signals (and their meanings) central to our analysis are preserved in the ALWD from the Bluebook (Dickerson, 1996).

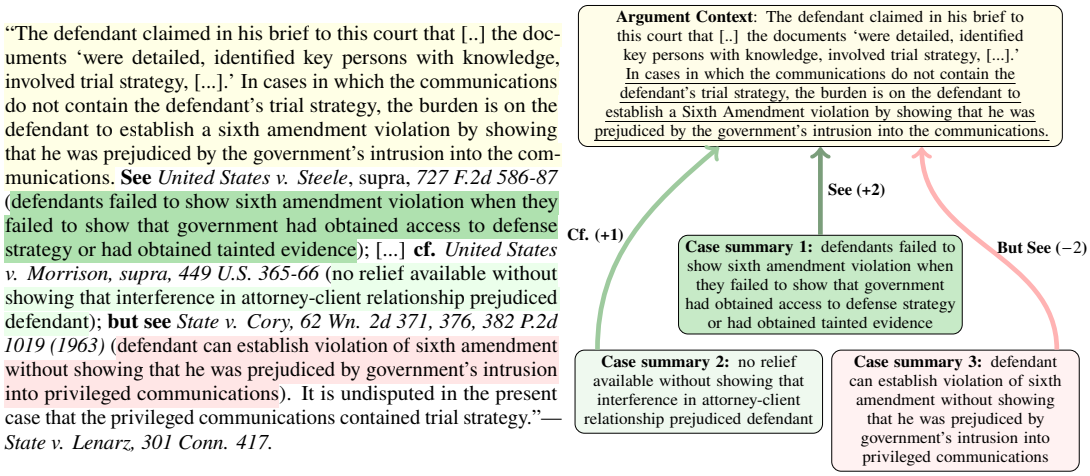

Figure 1: An example paragraph drawn from the judicial opinion, *State v. Lenarz, 301 Conn. 417*, showing a preceding argument context (with relevant text underlined for illustration purposes), and three relevant case citations. Each citation yields a stance-annotated tuple, extracted heuristically from the opinion based on the legal citation writing conventions, consisting of an argument context, a case summary extracted from the parenthetical associated with the cited case, and a stance value interpreted from the citation signal used in the original opinion.

> Given an *argument context* and a *case summary*, the task involves predicting the *stance value* by determining how the case summary modifies the argument context by evaluating along two dimensions: *polarity* (whether the case summary strengthens or weakens the argument context) and *intensity* (the degree of that modification).

Figure 1 shows an example from *State v Lenarz* on government intrusion of privileged attorney-client communications. The argument context discusses when defendants must prove harm from such intrusions to claim a violation, with three citations showing different stance values: *United States v. Steele* strongly supports the argument as both argue that the harm must always be proven, *United States v. Morrison* provides additional support though when seeking relief instead of violation claim, while *State v. Cory* opposes this requirement. [6]

Stance values. Prior research has established frameworks for analyzing semantic dimensions. For instance, Saurí and Pustejovsky (2012) analyzed event factuality using two dimensions: polarity (distinguishes between positive and negative instantiations of events) and modality (the degree of certainty, such as possible or probable). Similarly, Osgood (1957) considers two dimensions when studying semantic interpretations of adjectives: evaluation (word pairs like good-bad, beautiful-ugly, and happy-sad) and potency (word pairs like large-small, strong-weak, and heavy-light).

Drawing upon these prior semantic frameworks, we model our legal argument stances along two dimensions: *polarity* and *intensity*. We divide the *polarity* axis into positive (+) and negative (-), and the *intensity* axis is real-valued, capturing the level of support/opposition. A stance value is then characterized as a pair <polarity, intensity>, containing a polarity and intensity value (e.g., <+, 2>, abbreviated as +2). Finally, we add a neutral value on both axes to account for summaries that neither provide support nor opposition.

Figure 2 presents the full set of stance values, based on signal definitions prescribed in the *Bluebook*.[7] While the polarity can help highlight disagreements or contradictory cases for a given argument, intensity distinctions can help rank the cases such as identifying the most convincing case to build an argument (Gleize et al., 2019).

3.3 Tuple extraction process

We split judicial opinions into paragraphs using double-line breaks, then split each paragraph at the first citation signal. The text preceding the signal

[6] The original paragraph includes citations introduced with multiple citation signals. For illustration, we show citations cited via *see* and *cf.* signals to demonstrate varying intensity, and the *but see* signal to show different polarities. Other signals are omitted for visual clarity.

[7] The Bluebook has released several editions from 1926-present, each featuring a list of signals and their definitions. By the 1980s, these definitions largely stabilized into the form we present here. For a detailed comparison of definitions in different editions, see Dickerson (1996) and Appendix A.3: Table 4. For its impact on model performance, see §5.5.

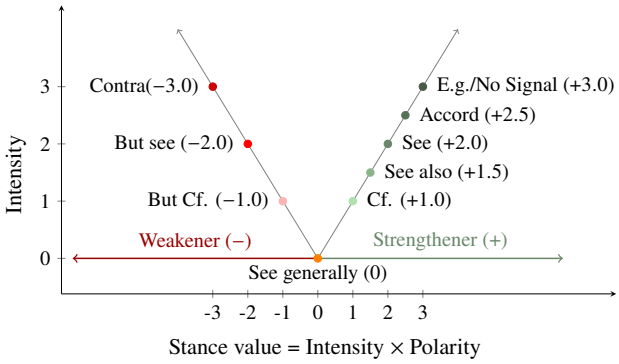

Figure 2: Stance values, ordered along two linguistic dimensions: *polarity* and *intensity*.

serves as argument context, excluding tuples with the context length ≤ 50 words. Manual inspection suggests this preceding text provides sufficient context to understand citation use.

We extract cited precedents and parenthetical texts from the text following the signal. To ensure each precedent correctly links to its relevant context, we restrict precedent extraction to the first sentence[8] after the signal—this avoids confusion when multiple arguments and precedents appear in the same paragraph. We then extract the case summary from the parenthetical text that appears directly after the citation.[9] To distinguish substantive explanations from citation metadata (dates, court information), we keep only spans >8 words.

Finally, we determine the stance values based on the definitions of the citation signals. We use the Bluebook's broad categorization of signals (support/opposition) to assign polarity and signal definitions to assign intensity values. *Cf* has the lowest intensity (+1) as it only "supports a proposition different from the main proposition but sufficiently analogous." *See also* ranks lower than *see* as it "supports the proposition, a bit less directly than a *see* cite." Signal pairs with opposite polarity but equivalent functions (e.g., *see* / *but see*; *cf* / *but cf*) receive equal intensity values with opposite signs. We assign the neutral value of 0 to *see generally*, defined as "cited authority provides background [..], without providing support for the conclusion."

Annotation quality. The stance labels in the extracted tuples are based on the citation signals assigned by actual US judges writing the legal opinions. As a result, our dataset contains high-quality professional annotations grounded in expert legal knowledge and normative linguistic rules, as defined in the Bluebook, which provides explicit guidance on the signal meaning followed by an expert community for a century. This approach also connects to established NLP traditions of dictionary-based semantics (e.g., Jurafsky and Martin (2025, Ch. 6) on lexical semantics, or word sense disambiguation as in Agirre and Rigau (1996); Yarowsky (1992)) and codebook-guided annotation (Halterman and Keith, 2024), also offering advantages over behavioral annotation-centric approaches that can face significant practical challenges in specialized domains such as law, including the requirement for highly specialized annotators and prohibitive costs of collecting such expert annotations.

4 Dataset Statistics

The average argument context length is 100 words (min:20, max:2545) while case summaries average 24 words (min:10, max:189). Support signals dominate, with the *see* signal being most frequent, followed by *see also* (Table 1), allowing authors to demonstrate well-supported arguments. Contradictory signals are less prevalent (*contra* rarest, then *but cf*), as practitioners often address contradictory cases substantively in text rather than just citing them with negative signals. Such negative signals are also more commonly found in academic writing, an interesting avenue to explore for future work. From a machine learning perspective, the significant class imbalance in stance labels presents challenges for multi-class classification, requiring balancing techniques (Japkowicz and Stephen, 2002), while the natural ordering of stances makes our dataset valuable for ordinal classification.

Signal	Stance Value	Total Occurrences
e.g.,	+3	244,545
accord	+2.5	33,697
see	+2	1,265,172
see also	+1.5	515,103
cf.	+1	152,786
see generally	0	23,962
but cf.	-1	5,371
but see	-2	26,954
contra	-3	1,772
All	-	2,269,362

Table 1: Number of tuples extracted per signal with corresponding stance values.

Temporal variation. The earliest tuple is from 1933, which is also when the Bluebook gained

[8]We built a sentence segmenter suited for legal texts, using a logistic regression model trained on lexical features (Gillick, 2009). Details are in Appendix A.2.

[9]We use *eyecite* (Cushman et al., 2021) to identify the location of a citation in a given text.

popularity after its release in 1926 (Cooper, 1982). A majority of tuples (74.68%) come from the opinions published between 2000-2020 (Figure 7, Appendix A.4).[10] As a result, our dataset primarily captures the contemporary citation practices followed in the legal community.

Most frequently cited cases Our dataset demonstrates strong face validity. The eight most frequently cited cases (cited>1000 times) establish key procedural standards governing how courts manage and decide cases. Examples include *Anderson v. Liberty Lobby*, *Celotex Corp. v. Catrett*, *Matsushita Electrical v. Zenith Radio* for summary judgments, and *Ashcroft v. Iqbal*, *Bell Atlantic v. Twombly* for standards to survive a motion to dismiss. Conversely, the cases receiving negative signals often did less well in establishing new standards. For example, *Solem v. Helm*, frequently cited negatively, has not been overruled, but narrowed by *Harmelin v. Michigan*, and now exists in legal limbo, illustrating how negative signals let judges acknowledge precedents, but also hold that it does not control their instant decision. For a detailed discussion, see Appendix A.5.

5 Experiments

5.1 Experimental Setup

Dataset splits. We use 14,283 tuples for training, 450 for validation, and 1,215 for testing,[11] with an equal number of tuples per class, thus ensuring that the model performance is not biased towards one class owing to class imbalance. We leave other class-balancing techniques for future work. We also ensure that the cases contributing to training tuples are disjoint from the cases for test tuples.

Models. Among proprietary LMs, we consider OpenAI's GPT-4o. We also evaluate BERT-base (fully open encoder-only transformer-based model) and Mistral-7B-Instruct (open-weight generative LM), along with their domain-pretrained variants LegalBERT (pre-trained on CAP corpus) and Saul-7B-Instruct (pre-trained on legal texts from multiple English-speaking jurisdictions (Colombo et al., 2024)), enabling analysis of domain-pretraining effects.

[10]This period also spans the Bluebook editions (17th-21st) during which citation signal definitions stayed consistent.

[11]A test set of 1,215 examples provides ±3% confidence intervals at 95% confidence level (Card et al., 2020).

Classification approaches. We explore *zero-shot prompting*, where we assume no labeled stance classification dataset is available. The goal is to assess existing models' performance without relying on large-scale data curation. We provide generative LMs with argument-summary pairs and signal definitions and ask them to classify the pair into one of the signals. We also explore *supervised fine-tuning* on δ-Stance. We take two approaches: (a) fine-tuning all parameters of the BERT-base (0.1B) model and (b) parameter-efficient fine-tuning of the Mistral-7B model using low-rank adaptation (Hu et al., 2021). Training details/prompts are provided in Appendix A.7.

Evaluation setup. For both zero-shot and supervised fine-tuning approaches, we develop a nine-class classifier and report F1 scores at three levels of granularity: nine, five, and three-class scale. This multi-granular evaluation helps identify model limitations and account for different use cases, from basic polarity to fine-grained stance classification.

5.2 Existing models' performance on the task.

Figure 3 compares model performance across citation signals. Fine-tuned models, FT-BERT-base(0.1B) and FT-Mistral-7B, outperform zero-shot ZS-GPT4o on most signals. While ZS-GPT4o performs well on coarse polarity distinctions (weakener F1: 0.77, strengthener F1: 0.81), it struggles with nine-class distinctions and distinguishing direct/indirect signals in the five-class setting (F1:0.35-0.39). Alternative prompting methods, like in-context learning and chain-of-thought (Kojima et al., 2022), show no significant improvements (Table 5 in Appendix A.6). Finally, during our experiments, two new models were released, o1 (OpenAI, 2024) and DeepSeek-R1 (DeepSeek-AI, 2025). Given their strong reasoning capabilities, we evaluated them in a zero-shot setting using the same prompt as GPT4o though found performance comparable to GPT4o (Table 6 in Appendix A.6). We leave a detailed analysis for future work.

Qualitative error analyses. We observe that zero-shot models often fail to process complex sentence structures prevalent in law. For instance, they often fail to distinguish between arguments from different legal actors (e.g., confusing a litigant's position with the position of the author of a judicial opinion). Models also overlook subordinate clauses in a case summary or argument context that can flip

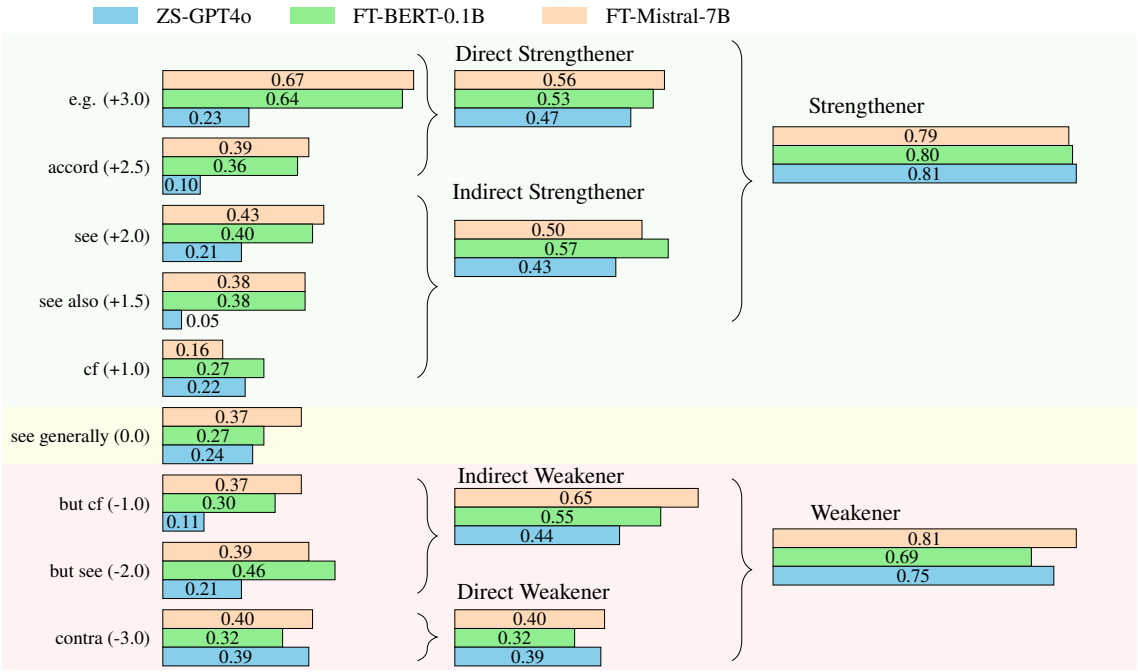

Figure 3: F1 scores of different stance classifiers at different granularity. ZS-GPT4o can distinguish polarity. Fine-tuning on δ-Stance helps improve fine-grained classification. ZS: zero-shot; FT: fine-tuned.

stance polarity. Fine-tuning helps mitigate some of these errors, as the model is explicitly trained on these complex sentence structures.

The fine-tuned models also perform comparably to the ZS-GPT4o on polarity classification, despite being much smaller in size. Among the different fine-grained classes, fine-tuning improves the performance most on the *e.g.,* class. In these tuples, the argument context often mentions a collective observation ('often courts have required [..]', 'A number of courts have concluded [..]') and the case summary provides an example supporting this observation (examples in Appendix A.8: Table 8). This pattern possibly results in linguistic cues that help the model identify the *e.g.,* signal. Finally, distinguishing the background class (*see generally*) remains challenging for both zero-shot and fine-tuned models, with the models mainly confusing the class with *see* and *e.g.,* classes. A detailed confusion matrix is provided in the Appendix A.6.

5.3 Can masked language models help predict stances in the zero-shot setting?

Our previous experiments relied on instructing generative LMs in the zero-shot setting. An alternative approach is to pre-train a masked language model (MLM) on CAP text and use it to assign probability to citation signal words given the surrounding context. Since MLMs are trained to identify words from the surrounding context, pre-training on legal text can expose the model to citation signal use in legal contexts, an approach used in prior work for developing context-sensitive language measures (Card et al., 2022; Cheng et al., 2024).

Method	Model	Domain	Base	Weighted F1
MLM	BERT-0.1B	General	-	0.13 ±0.04
	Legal-BERT-0.1B	Legal	BERT-0.1B	0.19 ±0.04
LLM Prompting	GPT4o	General	-	0.49 ±0.04
	Mistral-7B-Instruct	General	-	0.34 ±0.05
	Saul-7B-Instruct	Legal	Mistral-7B-Instruct	0.19 ±0.04
Supervised fine-tuning	BERT-0.1B	General	-	0.55 ±0.04
	Legal-BERT-0.1B	Legal	BERT-0.1B	0.54 ±0.04
	Mistral-7B-Instruct	General	-	0.58 ±0.05
	Saul-7B-Instruct	Legal	Mistral-7B-Instruct	0.59 ±0.04

Table 2: Stance classification performance comparing MLM, LLM prompting, and supervised fine-tuning approaches on a reduced 4-class (supported by MLM) version of the full task. Results show that supervised fine-tuning performs best, significantly outperforming both MLM and LLM methods. Colors visualize the proximity of that value to 1.0 (blue) and 0.0 (red).

We use LegalBERT, a BERT model pre-trained on CAP with domain-specific vocabulary. We concatenate each argument context and case summary pair with a [MASK] token and process this masked input through the model to compute the probabilities of the [MASK] token being filled by citation signals. We normalize these probabilities using softmax across all signals and select the signal with the highest normalized score. As MLMs are trained to predict one token at a time, we restrict this ex-

periment to four single-token citation signals in the model's vocabulary: *see*, *contra*, *accord*, and *cf*.

Table 2 shows that the MLM-based approach achieves very low F1 scores (F1: 0.13 – 0.19), lower than a random/majority baseline (F1: 0.25), despite being pre-trained on legal text. In contrast, `GPT-4o` demonstrates stronger zero-shot performance, achieving higher F1 scores across all citation signals. The models fine-tuned on δ-`Stance` consistently outperform MLM and prompting approaches, suggesting that MLM-based pre-training on legal text is insufficient for this task. The substantial performance gain from the supervised fine-tuning also highlights the value of δ-`Stance` in developing reliable stance classifiers.

5.4 Can domain pre-training help?

Prior work has demonstrated benefits of domain pre-training (e.g., to improve LMs' factuality (El Hamdani et al., 2024)). To examine whether pre-training improves stance classification, we compare domain-specific models against general-domain counterparts in zero-shot and supervised settings. For zero-shot setting, we use two strategies: a) *instructing generative LMs* (`Saul-LM-7B` and `Mistral-7B`) with prompts from §5.2, and b) using *masked language models* (`BERT` and `LegalBERT`) to predict citation signals from surrounding context. We also evaluate all four models after supervised fine-tuning. We consider four single-token signals for a fair comparison with the MLM approach.

As shown in Table 2, in a zero-shot setting, the domain-pretrained `LegalBERT` model performs comparably to the `BERT` model. Notably, domain pre-trained `Saul-LM-7B` obtains worse performance than the `Mistral-7B`, possibly owing to `Saul-LM`'s pre-training on texts from multiple jurisdictions with different legal traditions, affecting its performance on U.S.-specific δ-`Stance`. This observation suggests the importance of jurisdiction-specific pre-training, an interesting avenue for future work. Similarly, in the supervised setting, both model families show similar performance with and without domain pre-training. Overall, these observations suggest that the domain-pretraining remains insufficient and supervision from δ-`Stance` is more crucial.

5.5 Impact of temporal change in citation signal definitions on model performance

All our previous experiments assume consistent signal definitions across tuples. However, *Bluebook* definitions evolved significantly during early versions (Dickerson, 1996; Robbins, 1999), providing an opportunity to study how these changes affect model performance over time. Following prior work (Bamman, 2020; Yang and Eisenstein, 2015), we view this as a historical domain shift and evaluate our best-performing classifier, `FT-Mistral-7B`, on citation tuples from different time ranges.

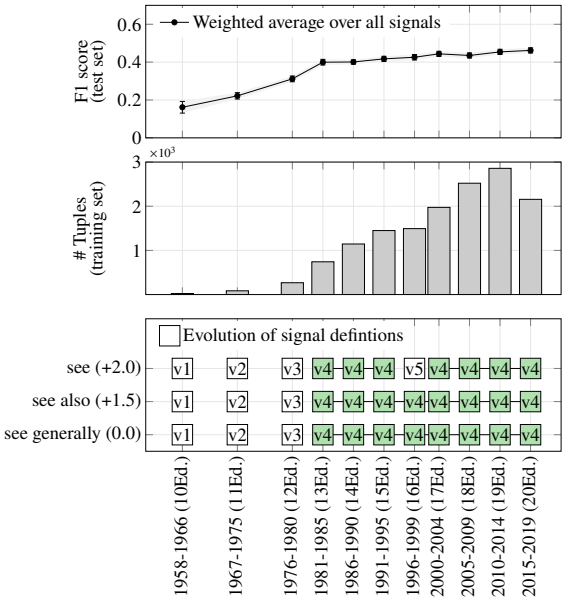

Figure 4: Fine-tuned Mistral-7B performance over the last 10 editions of the Bluebook. The model achieves lower F1 scores on historical tuples (10-13th editions) when the signal definitions changed frequently, followed by an improved performance in recent years.

Figure 4 shows the model performs poorly on historical citation tuples (10-13th Ed.) but improves in later years (14th-20th Ed.), suggesting it captures contemporary signal definitions. This effect is pronounced for signals with frequent definition changes (Figure 5), such as *see also* and *see generally*—from being equivalent (v2:11th Ed.), to indicating background material (v1:10th, v3:12th Ed.), to their current form (v4:13th Ed.-present). See Appendix A.3 for details. Furthermore, F1 scores remain stable after 1981 despite training tuple variations across editions, suggesting that once signal definitions stabilize, we can use tuples from data-rich years to classify tuples from years with fewer examples, as they share the same definitions.

Another notable change involves the *see* signal, whose definition was changed in the *Bluebook*'s 16th edition (v5) to indicate direct support (equivalent to no signal or *e.g.*), before reverting to its original meaning (v4) in the 17th edition (Dickerson, 1996; Robbins, 1999). However, model performance remained stable, suggesting the original

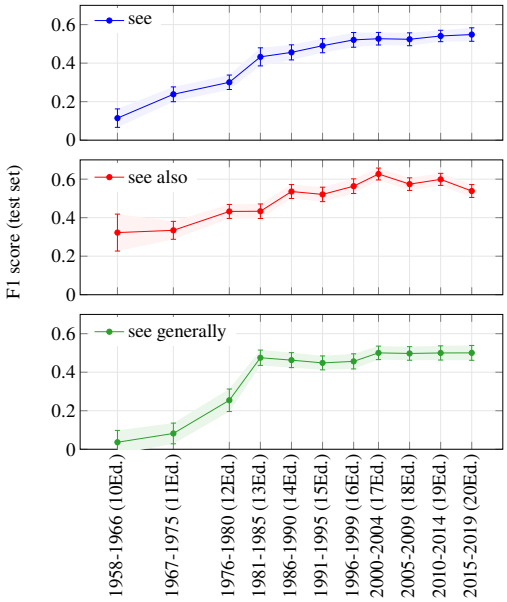

Figure 5: Fine-tuned Mistral-7B performance for individual signals over the last 10 editions of the Bluebook.

definition persisted in practice.

Overall, these observations highlight the challenges posed by evolving semantic conventions and the need for careful data curation based on downstream application requirements (Jo and Gebru, 2020). For instance, the classifier's reliance on contemporary signal definitions makes it well-suited for writing assistance tools but may limit its reliability for conducting historical analyses.

5.6 Generalization beyond the legal domain

We explore our dataset's utility beyond the law domain by testing if fine-tuning models on δ-Stance improves their performance on similar tasks in other domains. We evaluate models on the polarity classification task (whether a premise strengthens or weakens an argument). We train Mistral-7B as a binary classifier on δ-Stance, mapping positive signal tuples to strengtheners and negative to weakeners. Prompt details are in Appendix A.7.

We evaluate this classifier on two prior datasets: a) **ARCT** (Habernal and Gurevych, 2017): ARCT's test set contains 445 claim-reason pairs from news comments, each with two warrants - correct warrant that logically connects the reason to the claim and incorrect warrant that supports the claim's opposite. We treat claim-reason pairs as arguments, correct warrants as strengtheners, and incorrect warrants as weakeners, evaluating using weighted F1 scores. b) **ArguAna** (Wachsmuth et al., 2018): The ArguAna test set contains 1401 argument-counterargument pairs from debates spanning 15 domains (e.g., politics, science, health). For each argument, we treat counterarguments as weakeners and evaluate using recall (fraction of examples correctly identified as weakeners), as the dataset only contains weakeners.

Model	ARCT (Weighted F1)	ArguAna (Recall)
Mistral-7B (zero-shot)	0.69±0.03	0.76±0.02
Mistral-7B (FT on δ-Stance)	0.76±0.03	0.93±0.01
Mistral-7B (FT on mislabeled δ-Stance)	0.42±0.04	0.89±0.02
Mistral-7B (FT on target dataset)	0.86±0.02	0.98±0.02
GPT-4o (zero-shot)	0.82±0.03	0.95±0.01

Table 3: Model performance on arguments from different domains. The target dataset is ARCT/ArguAna.

Results. As shown in Table 3, fine-tuning on δ-Stance improves performance on both ARCT and ArguAna compared to the zero-shot setting. One explanation for the improved performance could be that fine-tuning simply enhances the model's instruction-following abilities (Hewitt et al., 2024). To investigate this, we train the model on a mislabeled version of δ-Stance, assigning a random label to an example. This model performs significantly worse on ARCT (0.42) and achieves 0.89 recall on ArguAna—better than the zero-shot model (0.76) but worse than fine-tuning on original δ-Stance (0.93)—suggesting that the model is learning meaningful relationships from δ-Stance. Finally, we also examine performance upper bounds using zero-shot GPT4-o prompting and direct fine-tuning on ARCT and ArguAna training data. While both methods outperform the δ-Stance fine-tuned model, they require either paid API calls or task-specific annotations. In contrast, our δ-Stance fine-tuned model offers reasonable performance without these resource requirements.

6 Conclusion

We present a large-scale dataset of stance-annotated legal argument pairs and evaluate several existing NLP methods on the *stance classification task* supported by this dataset. We also demonstrate the dataset's utility beyond law through improved cross-domain generalization, while also highlighting how changes in citation signal definitions can affect model performance. Several interesting analyses are possible as part of future work, such as using unsupervised factor analysis (Dodds et al., 2021) to study citation signal use in practice-based documents, or examining LLMs' influence on legal citation practices.

7 Limitations

We list some of the limitations of our study, which we hope will be useful for researchers and practitioners when interpreting our analysis.

1. First, while our dataset relies on citation signals for obtaining stance values, there may be an inherent subjectivity in how legal practitioners interpret and apply these signals. However, our stance labels originate from U.S. judges—arguably the highest-expertise annotators available for this task—who deeply research multiple cases, the current case being judged, and apply extensive legal training when making signal choices. Moreover, subjectivity in semantic annotation has also been reported in prior work (e.g., persistent low agreement among annotators for the NLI task (Pavlick and Kwiatkowski, 2019)), yet such datasets remain well-studied and useful for NLP, and similar argument relations are essential to legal discourse. Future work could explore methods to measure and incorporate this subjectivity in annotations (Plank, 2022).

2. Second, we interpret the normative definitions of each signal as prescribed in the *Bluebook*. However, the actual use of these signals by legal practitioners in judicial opinions may sometimes deviate from their normative definitions (e.g., owing to delay in adopting new definitions). Our analysis in §5.5 provides an example, where the change in *see* signal definition is not immediately reflected in the practice documents. Nevertheless, our dataset can help support such analyses in more detail, providing valuable insights into legal citation practices.

3. Third, our experiments evaluate several state-of-the-art language models, selecting representative examples from both proprietary and open-source model categories. While we provide these results, the landscape of available LLMs continues to evolve rapidly. Future work could explore newer models, including other proprietary, open-weight, and open-data models.

4. Fourth, our approach assumes the availability of case summaries as extracted from parentheticals. For writing assistance applications, such case summaries may not be always available. Legal case summarization is an active area of research (Akter et al., 2025) and interesting future work is possible to integrate a stance classification model with automatic summarization models.

5. Fifth, δ-Stance is currently limited to the U.S. legal citation system. However, many other jurisdictions (e.g., Australia and Canada) also employ similar citation signals in their legal writing, suggesting potential for broader application. Future work could explore using our approach to create datasets for other jurisdictions, further enabling cross-jurisdictional analysis of legal citation practices.

8 Ethics Statement

Our work is in line with the ACL Ethics Policy. The text and appendix outline all the models, datasets, and evaluation methodologies used in this research. The data used in this study is drawn from public textual data provided by CAP, containing legal opinions from the U.S. which are public records. No private or user-specific information beyond publicly accessible content was included. All other evaluation datasets used in this research are publicly available and used with the appropriate consent. The presented dataset is not intended to be a resource for anyone engaged in a legal dispute. δ-Stance is aimed to further practice-oriented legal NLP research and empirical legal studies and it also could form the basis for real-world systems that assist legal practitioners with their legal research.

9 Acknowledgements

We would like to thank the anonymous reviewers for their time and valuable feedback. We are grateful to Chloe Eggleston, Erica Cai, Marisa Hudspeth, Marzena Karpinska, Tessa Massis, and the UMass NLP group for several useful discussions during the course of the project. This material is based upon work supported by an IBM Ph.D. Fellowship awarded to AG. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily represent the official views of, or endorsement of, any sponsor.

References

Agnar Aamodt and Enric Plaza. 1994. Case-Based Reasoning: Foundational issues, methodological varia-

tions, and system approaches. *AI Communications*, 7(1):39–59.

Eneko Agirre and German Rigau. 1996. Word sense disambiguation using conceptual density. In *The 16th International Conference on Computational Linguistics*.

Mousumi Akter, Erion Cano, Erik Weber, Dennis Dobler, and Ivan Habernal. 2025. A comprehensive survey on legal summarization: Challenges and future directions. *ArXiv*, abs/2501.17830.

Kevin D Ashley and Edwina L Rissland. 1987. But, See, Accord: Generating "blue book" citations in HYPO. In *Proceedings of the 1st International Conference on Artificial Intelligence and Law*, pages 67–74.

David Bamman. 2020. *LitBank: Born-Literary Natural Language Processing*. Debates in the Digital Humanities.

Douglas Biber and Edward Finegan. 1989. Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Text - Interdisciplinary Journal for the Study of Discourse*, 9(1):93–124.

Leonard Bongard, Lena Held, and Ivan Habernal. 2022. The legal argument reasoning task in civil procedure. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 194–207, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Dallas Card, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Platt Boustan, Ran Abramitzky, and Dan Jurafsky. 2022. Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration. *Proceedings of the National Academy of Sciences of the United States of America*, 119.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Conference on Empirical Methods in Natural Language Processing*.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. 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, Dublin, Ireland. Association for Computational Linguistics.

Jennifer Elisa Chapman. 2024. Teaching critical use of legal research technology. *Legal Writing Journa*, 28(1).

Myra Cheng, Kristina Gligoric, Tiziano Piccardi, and Dan Jurafsky. 2024. AnthroScore: A computational linguistic measure of anthropomorphism. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 807–825.

Odysseas S. Chlapanis, Dimitrios Galanis, and Ion Androutsopoulos. 2024. LAR-ECHR: A new legal argument reasoning task and dataset for cases of the European court of human rights. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 267–279, Miami, FL, USA. Association for Computational Linguistics.

Pierre Colombo, Telmo Pires, Malik Boudiaf, Rui Melo, Dominic Culver, Etienne Malaboeuf, Gabriel Hautreux, Johanne Charpentier, and Michael Desa. 2024. SaulLM-54B & SaulLM-141B: Scaling up domain adaptation for the legal domain. In *Advances in Neural Information Processing Systems*, volume 37, pages 129672–129695. Curran Associates, Inc.

Byron D. Cooper. 1982. Anglo-American legal citation: Historical development and library implications. *75 Law Library Journal 3*.

Jack Cushman, Matthew Dahl, and Michael Lissner. 2021. eyecite: A tool for parsing legal citations. *Journal of Open Source Software*, 6(66):3617.

Xiang Dai, Ilias Chalkidis, Sune Darkner, and Desmond Elliott. 2022. Revisiting transformer-based models for long document classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7212–7230, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rajarshi Das, Manzil Zaheer, Dung Thai, Ameya Godbole, Ethan Perez, Jay Yoon Lee, Lizhen Tan, Lazaros Polymenakos, and Andrew McCallum. 2021. Case-based reasoning for natural language queries over knowledge bases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9594–9611, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Preprint*, arXiv:2501.12948.

Darby Dickerson. 1996. An Un-Uniform System of Citation: Surviving with the New Bluebook. 26 Stetson L. Rev. 53.

Peter Sheridan Dodds, Thayer Alshaabi, Mikaela Irene D. Fudolig, J. W. Zimmerman, Jerome Lovato, Samuel Beaulieu, Joshua R. Minot, Michael V. Arnold, Andrew J. Reagan, and Christopher M. Danforth. 2021. Ousiometrics and telegnomics: The essence of meaning conforms to a two-dimensional powerful-weak and dangerous-safe framework with diverse corpora presenting a safety bias. *ArXiv*, abs/2110.06847.

Jesse Dodge, Ana Marasović, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In *Conference on Empirical Methods in Natural Language Processing*.

Rajaa El Hamdani, Thomas Bonald, Fragkiskos D. Malliaros, Nils Holzenberger, and Fabian Suchanek. 2024. The factuality of large language models in the legal domain. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 3741–3746, New York, NY, USA. Association for Computing Machinery.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. 2024. LawBench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7933–7962, Miami, Florida, USA. Association for Computational Linguistics.

Ian Gallacher. 2006. Cite Unseen: How neutral citation and America's law schools can cure our strange devotion to bibliographical orthodoxy and the constriction of open and equal access to the law. Available at SSRN.

Daniel Gillick. 2009. Sentence boundary detection and the problem with the U.S. In *North American Chapter of the Association for Computational Linguistics*.

Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2019. Are you convinced? Choosing the more convincing evidence with a Siamese network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 967–976, Florence, Italy. Association for Computational Linguistics.

Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. 2023. Summary of the competition on legal information, extraction/entailment (COLIEE) 2023. *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*.

Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. 2024. Overview of benchmark datasets and methods for the legal information extraction/entailment competition (COLIEE) 2024. In *New Frontiers in Artificial Intelligence*, pages 109–124, Singapore. Springer Nature Singapore.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2024. 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.

Ivan Habernal, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Indra Spiecker genannt Döhmann, and Christoph Burchard. 2024. Mining legal arguments in court decisions. *Artificial Intelligence and Law*, pages 1–38.

Ivan Habernal and Iryna Gurevych. 2016. What makes a convincing argument? Empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223, Austin, Texas. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Andrew Halterman and Katherine A. Keith. 2024. Codebook LLMs: Evaluating LLMs as measurement tools for political science concepts. volume abs/2407.10747.

John Hewitt, Nelson F. Liu, Percy Liang, and Christopher D. Manning. 2024. Instruction following without instruction tuning. *ArXiv*, abs/2409.14254.

Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2020. A dataset for statutory reasoning in tax law entailment and question answering. In *NLLP@KDD*.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685.

Nathalie Japkowicz and Shaju Stephen. 2002. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449.

Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, page 306–316, New York, NY, USA. Association for Computing Machinery.

Abhinav Joshi, Shounak Paul, Akshat Sharma, Pawan Goyal, Saptarshi Ghosh, and Ashutosh Modi. 2024. IL-TUR: Benchmark for Indian legal text understanding and reasoning. In *Annual Meeting of the Association for Computational Linguistics*.

Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released January 12, 2025.

Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael James Bommarito. 2023. Natural language processing in the legal domain. *SSRN Electronic Journal*. Available at SSRN.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Janet Kolodner. 1992. An introduction to Case-Based Reasoning. *Artificial Intelligence Review*, 6:3–34.

Robert Koons. 2022. Defeasible Reasoning. In *The Stanford Encyclopedia of Philosophy*, Summer 2022 edition. Metaphysics Research Lab, Stanford University.

William M. Landes, Lawrence Lessig, and Michael E. Solimine. 1998. Judicial influence: A citation analysis of federal courts of appeals judges. *The Journal of Legal Studies*, 27(2):271–332.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45:765–818.

Zehua Li, Neel Guha, and Julian Nyarko. 2023. Don't use a cannon to kill a fly: An efficient cascading pipeline for long documents. *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Kelvin Luu, Xinyi Wu, Rik Koncel-Kedziorski, Kyle Lo, Isabel Cachola, and Noah A. Smith. 2021. Explaining relationships between scientific documents. 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 2130–2144, Online. Association for Computational Linguistics.

Robert Mahari, Dominik Stammbach, Elliott Ash, and Alex 'Sandy' Pentland. 2024. LePaRD: A large-scale dataset of judicial citations to precedent. In *Annual Meeting of the Association for Computational Linguistics*.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Mitra Mohtarami, Ramy Baly, James Glass, Preslav Nakov, Lluís Màrquez, and Alessandro Moschitti. 2018. Automatic stance detection using end-to-end memory networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 767–776, New Orleans, Louisiana. Association for Computational Linguistics.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Sturmer, and Ilias Chalkidis. 2023. LEXTREME: A multi-lingual and multi-task benchmark for the legal domain. In *Conference on Empirical Methods in Natural Language Processing*.

OpenAI. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Charles E Osgood. 1957. The measurement of meaning. *Urbana: University of Illinois Press*.

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*, page 98–107, New York, NY, USA. Association for Computing Machinery.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Eitan Adam Pechenick, Christopher M Danforth, and Peter Sheridan Dodds. 2015. Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PloS one*, 10(10):e0137041.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Prakash Poudyal, Jaromír Šavelka, Aagje Ieven, Marie-Francine Moens, Teresa Gonçalves, and Paulo Quaresma. 2020. ECHR: Legal corpus for argument mining. In *Workshop on Argument Mining*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Conference on Empirical Methods in Natural Language Processing*.

Ira P Robbins. 1999. Semiotics, Analogical Legal Reasoning, and the Cf. Citation: Getting our signals uncrossed. *Duke Law Journal*, 48:1043–1080.

Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675, Online. Association for Computational Linguistics.

Linda M Ryan. 2022. The ALWD guide to legal citation: One small step toward wellness in legal education.

George Sanchez. 2019. Sentence boundary detection in legal text. *Proceedings of the Natural Legal Language Processing Workshop 2019*.

Roser Saurí and James Pustejovsky. 2012. Are you sure that this happened? Assessing the factuality degree of events in text. *Computational Linguistics*, 38(2):261–299.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551.

Benjamin Schmidt, Steven T. Piantadosi, and Kyle Mahowald. 2021. Uncontrolled corpus composition drives an apparent surge in cognitive distortions. *Proceedings of the National Academy of Sciences*, 118(45):e2115010118.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Annual Meeting of the Association for Computational Linguistics*.

Benjamin Weiser. 2023. ChatGPT lawyers are ordered to consider seeking forgiveness. *New York Times*.

Yi Yang and Jacob Eisenstein. 2015. Unsupervised multi-domain adaptation with feature embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 672–682, Denver, Colorado. Association for Computational Linguistics.

David Yarowsky. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *The 14th International Conference on Computational Linguistics*.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2016. Ordinal common-sense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.

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*, page 159–168.

A Appendix

A.1 Example tuples from our dataset

The earliest stance tuple in our dataset is from a judicial opinion published in 1933.

> One may tear down a building and deliver the lumber and other material, or sever the building and sell it in its entirety and thus sell personal property, or an owner may permit a building to be placed on his land, agree that it is not to be affixed, and permit the owner to come upon the premises and remove the building. **See** *Clements v. Morton, 200 Ala. 390, 76 So. 306*; (where a house was erected on land of another with the understanding it was to be the property of the one building it). *Rogers v. Cox, 96 Ind. 157, 49 Am. Rep. 152* (where it is said there was no proof that the building was ever annexed to the land). Brown v. Roland, 11 Tex. Civ. App. 648, 33 S. W. 273.'—*Baird v. Elliott, 63 N.D. 738.*

A.2 Sentence segmenter details

Following the approach of Gillick (2009), for a given opinion text, we identify potential sentence boundaries by extracting all text spans of the form "L. R," where L is the word on the left side of the period in question, and R is the word on the right. We then predict the probability of the binary sentence boundary class s, conditional on its context: $P(s|\text{“}L.\ R\text{''})$, where we use a logistic regression model to predict probability based on features extracted from "L. R" (e.g., whether L is capitalized or both L and R are lowercase). We use all the features from Gillick (2009), along with our custom features specific to legal texts (e.g., whether R is a year, L is *v.*), listed in Table X.

For training, we use the sentence boundary annotations provided in Sanchez (2019). First, we extract all text spans of the form "L. R". Spans that occur within a single sentence are labeled as non-boundaries, while spans that crossed sentence boundaries (i.e., where L was the last word of one sentence and R was the first word of the next) are labeled as boundaries. We train and test the logistic regression model on all the annotated text spans. Our model achieves a 90% weighted F1 on the sentence boundary prediction task across all text spans extracted from sentences in the test set. Finally, we

Edition	Publication year	Changes to citation signals
1st	1926	First release; mentions *see*, *contra*, *cf*, and *but see* signals
2nd	1928	Introduced *but cf*
3rd	1931	No changes
4th	1934	No changes
5th	1936	Introduced *accord*
6th	1939	No changes
7th	1947	Introduced *e.g.,* signal; clarified definition of other signals
8th	1949	Reorganized and clarified definitions of all signals
9th	1955	Clarified definitions of all signals
10th	1958	Introduced *see also* and *see generally* as separate supplementary signals
11th	1967	Unified *see generally* and *see also* as equivalent
12th	1976	Introduced *see also* and *see generally* as separate supplementary
13th	1981	Redefined *see also* as support signal
14th	1986	No changes
15th	1991	No changes
16th	1996	Revised definition of *see* signal; removed *contra* signal
17th	2000	Reverted to 15th edition's definitions
18th	2005	No changes
19th	2010	No changes
20th	2015	No changes
21st	2020	No changes; current version

Table 4: Historical evolution of The Bluebook citation signals (1926-2020). The table tracks major changes to the definitions of citation signals.

use the predictions for each text span to segment a given opinion text into sentences, splitting the opinion at each text span predicted as a sentence boundary.

A.3 Different editions of the Bluebook

Table 4 provides an overview of how signal definitions changed over time.

Figure 6 further demonstrates the performance of our stance classifier over tuples from different time periods. Additionally, it shows the semantic similarity of a signal definition w.r.t the modern definition (21st edition). For all three signals, we find a strong positive correlation between the F1 score and the semantic similarity[12] between a signal's definition and its modern definition (21st Ed.) (Spearman's ρ: 0.80 (*see*), 0.72 (*see also*), 0.78 (*see generally*)) more details in Appendix A.3), suggesting that model performance increases as signal definitions evolve closer to their modern forms.

A.4 Temporal stance distribution

Figure 7 shows the temporal distribution of tuples for each signal.

[12]using all-mpnet-base-v2 (Reimers and Gurevych, 2019).

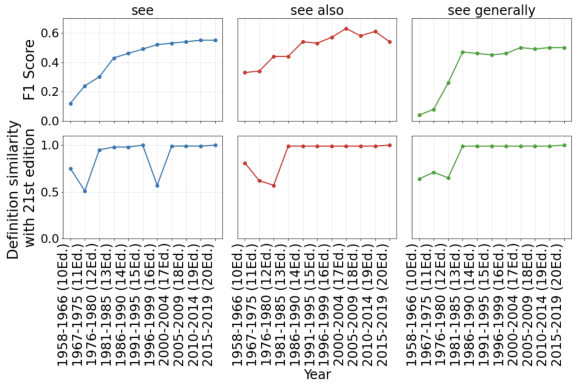

Figure 6: Stance classifier performance along with variation in signal definitions over time.

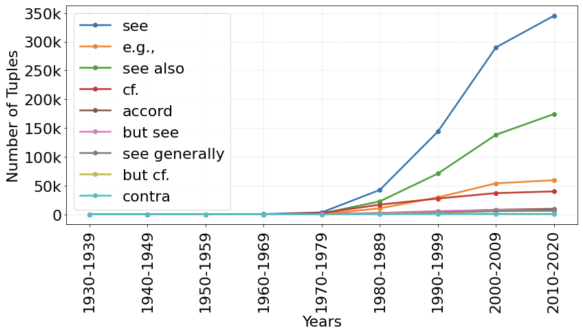

Figure 7: Number of stance tuples extracted per decade.

A.5 Most frequently cited cases in δ-Stance dataset.

Figure 8 shows the most frequently cited cases in our dataset, along with how often these cases were cited using different citation signals.

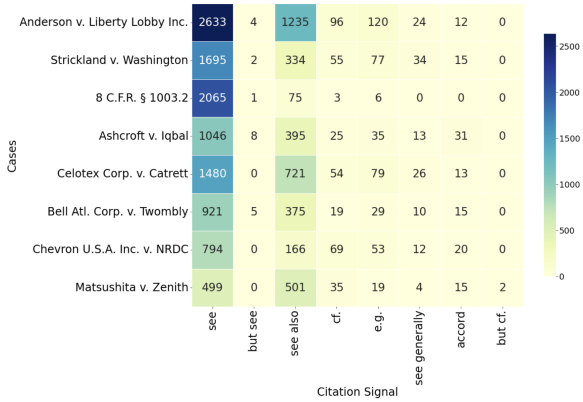

Figure 8: Most frequently cited cases in our dataset and their stance distribution using the gold standard labels.

The eight most frequently cited cases in our dataset (cited>1000 times) establish key procedural standards–rules that govern how courts manage and decide cases. For instance, the U.S. Supreme Court case *Anderson v. Liberty Lobby*, which articulates the standard regarding motions for summary judgment, is the most cited and often cited with

supporting signals (e.g., see, see also)–unsurprising given summary judgment is the most common procedural mechanisms frequently used in determining whether a case will proceed to trial. Other oft-cited cases are likewise procedural: *Ashcroft v. Iqbal* and *Bell Atlantic v. Twombly* deal with standards for civil cases to survive a motion to dismiss; *Celotex Corp. v. Catrett* specify burden of proof in summary judgment motions and interpretation of a key part of Federal Rules of Civil Procedure, and *Matsushita Electrical v. Zenith Radio* specify standards for summary judgment in antitrust cases.

On the other hand, cases frequently receiving negative signals (though at lower magnitudes) include *Floral Trade Council v. U.S.* due to its complexity and the regulatory ambiguity, *Wellness International Network v. Sharif* due to the defendant's behavior and the limits of non-Article III courts' authority, and *Solem v. Helm* (463 U.S. 277). *Solem v. Helm* is particularly interesting - while not overruled, its three-part test for determining disproportionate criminal punishments potentially violating the 8th Amendment was later narrowed by *Harmelin v. Michigan* and now exists in legal limbo. leading judges to acknowledge but not follow it. Therefore, the treatment with more negative signals operates as a mechanism for judges to acknowledge a case's existence, but to hold that it does not control their instant decision.

Overall, our dataset demonstrates strong face validity, with the most frequently cited cases establishing key procedural standards, reflecting their critical role in shaping how courts manage and decide cases. As such, the dataset demonstrates the enduring influence of these cases on procedural decisions in the U.S. courts, and the key role of procedure towards judicial decision-making. Conversely, cases that did less well in establishing new standards provide evidence that our approach can also uncover the flip side of the coin, with precedents that are difficult to apply leading to more negative interactions in future decisions.

A.6 Detailed performance of models

Confusion matrix for nine-class classifiers. Figure 9 shows the confusion matrix for three models: GPT4o in a zero-shot setting, fine-tuned BERT, and fine-tuned Mistral-7B model.

Alternative prompting approaches for GPT4-based nine-class classifier. Table 5 shows per-class F1 scores across different prompting approaches.

Signals	ZS-GPT4o	CoT-GPT4o	9Shot-GPT4o	FT-BERT-0.1B	FT-Mistral-7B
e.g., (+3.0)	0.23±0.05	0.24±0.04	0.2±0.06	0.64±0.07	0.67±0.06
accord (+2.5)	0.10±0.06	0.04±0.04	0.12±0.07	0.36±0.07	0.39±0.08
see (+2.0)	0.21±0.06	0.13±0.06	0.22±0.06	0.40±0.07	0.43±0.07
see also (+1.5)	0.05±0.05	0.08±0.06	0.15±0.05	0.38±0.07	0.38±0.07
cf (+1.0)	0.22±0.07	0.24±0.07	0.30±0.08	0.27±0.07	0.16±0.07
see generally (0.0)	0.24±0.08	0.19±0.07	0.3±0.09	0.27±0.08	0.37±0.07
but cf (-1.0)	0.11±0.06	0.22±0.08	0.25±0.08	0.3±0.08	0.37±0.07
but see (-2.0)	0.21±0.06	0.18±0.07	0.27±0.07	0.46±0.07	0.39±0.07
contra (-3.0)	0.39±0.07	0.41±0.07	0.35±0.08	0.32±0.08	0.4±0.08
weighted average	0.20±0.02	0.19±0.02	0.24±0.02	0.38±0.03	0.40±0.03

Table 5: F1 scores across different models and prompting strategies. ZS: zero-shot; CoT: chain-of-thought; 9Shot: in-context learning with one demonstration per class; FT: fine-tuned.

Signals	ZS-GPT4o	OpenAI o1	DeepSeek-R1
e.g., (+3.0)	0.23±0.05	0.27±0.04	0.26±0.04
accord (+2.5)	0.10±0.06	0.26±0.07	0.33±0.07
see (+2.0)	0.21±0.06	0.17±0.06	0.06±0.05
see also (+1.5)	0.05±0.05	0.01±0.02	0.04±0.04
cf (+1.0)	0.22±0.07	0.23±0.07	0.23±0.07
see generally (0.0)	0.24±0.08	0.08±0.06	0.13±0.07
but cf (-1.0)	0.11±0.06	0.08±0.06	0.18±0.08
but see (-2.0)	0.21±0.06	0.18±0.07	0.21±0.07
contra (-3.0)	0.39±0.07	0.50±0.07	0.43±0.07
weighted average	0.20±0.02	0.20±0.02	0.21±0.02

Table 6: Performance comparison measured using F1 scores across different models.

Detailed performance of four-class classifiers. Table 7 shows per-class F1 scores across different classification approaches.

A.7 Model training and prompt details

We fine-tune `BERT-0.1B` for 10 epochs, with a learning rate of $5e^{-6}$ and a batch size of 32, selecting the checkpoint with the lowest validation loss. For `Mistral-7B`, we fine-tune low-rank adapter matrices in attention and feed-forward layers for 1 epoch with a learning rate of $2e^{-4}$ and batch size of 8. Fig-

Method	Model	F1				
		contra (-3.0)	cf (+1.0)	see (+2.0)	accord (+2.5)	weighted
MLM	BERT	0.06±0.05	0.01±0.03	0.41±0.05	0.06±0.05	0.13 ±0.04
	Legal-BERT	0.04±0.05	0.07±0.05	0.42±0.05	0.23±0.08	0.19 ±0.04
LLM Prompting	GPT4o	0.68±0.06	0.39±0.07	0.49±0.06	0.38±0.08	0.49 ±0.04
	Mistral-7B-Instruct	0.48±0.09	0.40±0.08	0.43±0.07	0.06±0.06	0.34 ±0.05
	Saul-7B-Instruct	0.01±0.03	0.39±0.05	0.19±0.07	0.15±0.07	0.19 ±0.04
Supervised fine-tuning	BERT	0.54±0.08	0.38±0.08	0.63±0.06	0.64±0.06	0.55 ±0.04
	Legal-BERT	0.50±0.07	0.41±0.06	0.61±0.09	0.62±0.06	0.54 ±0.04
	Mistral-7B-Instruct	0.75±0.05	0.17±0.08	0.70±0.06	0.71±0.06	0.58 ±0.05
	Saul-7B-Instruct	0.74±0.05	0.25±0.05	0.69±0.08	0.69±0.07	0.59 ±0.04
Baselines	Random (uniform)	0.25±0.06	0.25±0.06	0.25±0.06	0.25±0.06	0.25 ±0.03
	Majority class	0.00	0.00	0.40	0.00	0.25

Table 7: Stance classification performance comparing zero-shot methods (using MLM and LLM prompting), supervised fine-tuning, and baseline models on four single-token citation signals, supported by the MLM framework. Colors visualize the proximity of that value to 1.0 (blue) and 0.0 (red).

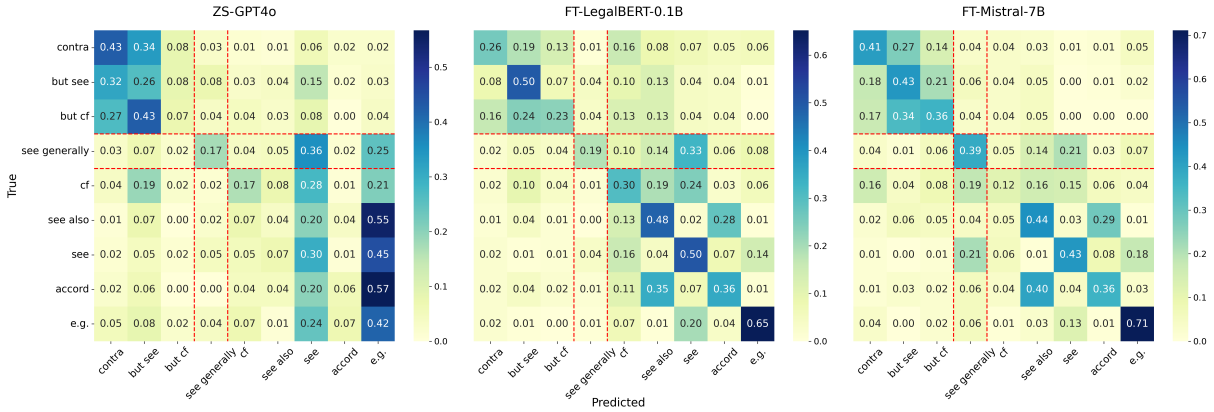

Figure 9: Confusion matrix for GPT4o and the fine-tuned models.

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction: You will be provided with an argument and a premise. Your task is to determine the most appropriate citation signal to connect the premise with the argument.
Answer choices:
(1) Contra
(2) Cf
(3) See
(4) Accord

Input:
Argument: <argument>{}</argument>
Premise: <premise>{}</premise>

Response: {}

Figure 10: Prompt used for fine-tuning Mistral-7B model for four-class stance classification.

ure 10, Figure 12 and Figure 11 show the prompts used.

A.8 Qualitative Analysis

Table 8 provides example tuples with 'e.g.' as the gold label.

Suppose you are a legal practitioner. You will be shown a legal argument and a premise. A premise can be connected to the argument using one of the following citation signals.

Contra: Indicates that the premise directly states the opposite of whatever is said in the legal argument.
But see: Indicates that the premise clearly supports the opposite of whatever is said in the argument, although an inference is required to see the contradiction.
But cf.: Indicates that the premise supports something similar to the opposite of whatever is said in the legal argument.
E.g.,: Indicates that the premise directly supports the legal argument.
See: Indicates that the premise clearly, though indirectly, supports the legal argument.
Accord: It indicates that two or more sources state or support the legal argument, but the argument either quotes or refers to only one of them. It is most often used to indicate that the law of one jurisdiction is in line with that of another. When using accord, your first introduce the source and its premise that the legal argument quotes or refers to, generally using [no signal], and then use accord to introduce the source and its premise that was not directly referred to.
See also: Indicates that the premise supports the legal argument, albeit a bit less directly or a bit less forcefully than a premise cited using the "see" signal.
Cf.: This signal is used to introduce a premise that supports a legal argument that is different from the one it follows, but that is analogous enough to the argument that it still indirectly supports the legal argument.
See generally: This signal is used to provide readers with information that they may refer to in order to better understand the background of the legal argument.

Your task is to identify the appropriate citation signal to connect the given premise with the given argument.
Argument: {}
Premise: {}

Question: What is the appropriate citation signal to connect the given premise with the given argument? If you can't tell what it is, say "Could not classify." Provide your reasoning first and then the final answer in the next line. The final answer should be only the citation signal name and nothing else."

Figure 11: Prompt for zero-shot nine class stance classification used for prompting `GPT4o`, `OpenAI o1` and `DeepSeek-R1` models.

Argument Context	Case Summary
"[R]elief under Section 11(a) [of the FAA) is limited to 'simple formal, descriptive, or mathematical mistake,' Stroh Container Co. v. Delphi Industries, Inc., 783 F.2d 743, 749 (8th Cir. 1986), not disagreement over factual or legal decisions deliberately made. Most cases discussing Section 11(a) address the alleged miscalculation of figures. **These cases make clear** that the provision reaches only computational errors, not legal or factual mistakes concerning the amount of damages that should be awarded.	error in determining start and stop dates for interest is not correctable under Section 11(a).
The district court's order did not give similar attention to Mr. Ellibee's remaining claims. In particular, we note the lack of discussion related to allegations that the parole board retaliated against him because of his litigation activities on behalf of himself and other prisoners. "Prison officials may not retaliate against ... an inmate because of the inmate's exercise of his right of access to the courts" and "[i]t is well established that prison officials may not unreasonably hamper inmates in gaining access to the courts." Smith v. Maschner, 899 F.2d 940, 947 (10th Cir.1990). **This court and other federal courts have recognized** actionable constitutional claims in inmates' allegations of denial of parole in retaliation for filing lawsuits.	reversing a district court's dismissal of a parole-retaliation claim as frivolous.
Regarding the first factor, to determine if a natural father of a newborn child has taken diligent, affirmative action, courts measure the putative father's efforts to make a financial commitment to the upbringing of the child, to legally substantiate his relationship with the child, and to provide emotional, financial, and other support to the mother during the pregnancy. Following the holdings in Quilloin and Lehr, **often courts have required** the father to use those legal mechanisms within his control that would entitle him to notice under the state's statutes, i.e., acknowledge or prove paternity, agree to a support order, or file with a putative father registry, and have done so even if a statute does not specify that adherence is required.	no constitutional infirmity in state court proceedings in which biological father was excluded because he had failed to establish parentage according to state law.
We stress that there are many different types of situations in which multiplicity issues arise, and we do not purport to set forth a general waiver rule covering all situations. At the same time, we note that **this is not the first time** we have found a double jeopardy claim waived.	failure to raise multiplicity claim resulted in treating the claim in the ineffective assistance context after trial.
The availability of interest in an action arising under a federal statute is governed by federal law, not the law of the forum state. See Norfolk & Western Railway Co. v. Liepelt, 444 U.S. 490, 493, 100 S.Ct. 755, 757, 62 L.Ed.2d 689 (1980) ("questions concerning the measure of damages in an FELA action are federal in character"); Faulkenberry v. Louisiana & Arkansas Railway Co., 551 F.2d 650 (5th Cir.1977). Title 28 U.S.C. § 1961 provides for postjudgment interest on money damages recovered in federal court. Neither this section nor the FELA itself, however, contains any provision concerning the availability of prejudgment interest as part of a plaintiffs compensation. **A number of courts have concluded** on the basis of Congress' silence that state laws authorizing prejudgment interest cannot be invoked in an FELA action.	Congress' silence concerning prejudgment interest is "indicative of a considered purpose that no interest should be allowed in [FELA] actions prior to verdict;" state statutes are therefore superseded

Table 8: Examples demonstrating a pattern in tuples with the *e.g.,* gold label, where the argument context presents a broad observation about multiple cases or instances, and the case summary provides a specific supporting example.

```
Below is an instruction that describes a task, paired with an input that provides
further context. Write a response that appropriately completes the request.

### Instruction: You will be provided with an argument and a premise.
Your task is to determine the most appropriate citation signal to connect the
premise with the argument.
Answer choices:
(1) Contra
(2) But see
(3) But cf
(4) See generally
(5) Cf
(6) See also
(7) See
(8) Accord
(9) E.g.,


### Input:
Argument: <argument>{}</argument>
Premise: <premise>{}</premise>


### Response: {}
```

Figure 12: Prompt used for fine-tuning Mistral-7B model for nine-class stance classification.