Received: 15 March 2026 | Revised: 4 June 2026 | Accepted: 30 June 2026 | Published online: 28 July 2026

RESEARCH ARTICLE

Journal of Computational Law and Legal Technology
2026, Vol. 00(00) 1–19
DOI: 10.47852/bonviewJCLLT62029649

BON VIEW PUBLISHING

LAMUS: A Large-Scale Corpus for Legal Argument Mining from U.S. Caselaw Using LLMs

Serene Wang[1], Lavanya Pobbathi[2] and Haihua Chen[2,*]

[1] *Texas Academy of Mathematics and Science, University of North Texas, USA*

[2] *College of Information, University of North Texas, USA*

Abstract: Legal argument mining aims to identify and classify the functional components of judicial reasoning, such as facts, issues, rules, analysis, and conclusions. Progress in this area is limited by the lack of large-scale, high-quality annotated datasets for U.S. caselaw, particularly at the state level. This paper introduces LAMUS, a sentence-level legal argument mining corpus constructed from U.S. Supreme Court decisions and Texas criminal appellate opinions. The dataset is created using a data-centric pipeline that combines large-scale case collection, large language model (LLM)-based automatic annotation, and targeted human-in-the-loop quality refinement. We formulate legal argument mining as a six-class sentence classification task and evaluate multiple general-purpose and legal-domain language models under zero-shot, few-shot, and chain-of-thought prompting strategies, with LegalBERT as a supervised baseline. Results show that chain-of-thought prompting substantially improves LLM performance, while domain-specific models exhibit more stable zero-shot behavior. LLM-assisted verification corrects nearly 20% of annotation errors, improving label consistency. Human verification achieves Cohen's kappa $\kappa = 0.85$, confirming annotation quality. LAMUS provides a scalable resource and empirical insights for future legal natural language processing research. All code and datasets can be accessed for reproducibility in GitHub https://github.com/LavanyaPobbathi/LAMUS.

Keywords: legal argument mining, legal argument extraction, Supreme Court Caselaw, corpus construction, large language models

1. Introduction

Argument mining aims to automatically identify and structure argumentative components in text, enabling downstream reasoning and decision-support tasks [1, 2]. In the legal domain, argument mining encompasses the extraction and classification of facts, issues, rules, legal analysis, and conclusions, which together form the basis of judicial reasoning. Automating these processes has the potential to significantly improve legal research efficiency, support judicial decision-making, and advance AI-driven legal intelligence [3, 4].

Legal argument mining (LAM) presents unique challenges due to specialized legal language, hierarchical reasoning structures, and extensive reliance on precedent and statutory interpretation [5, 6]. Recent advances in large language models (LLMs) have increased interest in applying these systems to legal reasoning and argument analysis tasks [7–9]. Judicial opinions rely on specialized language, hierarchical reasoning, and extensive references to precedent and statute, making both annotation and modeling substantially more complex than in general-domain argument mining. Typical workflows involve argument component detection, argument classification, and, in more advanced settings, reconstruction of argumentative structures and reasoning chains. The rapid development of LLMs has led to increasing interest in evaluating their performance in specialized legal tasks and workflows [7, 8].

Recent research has advanced fine-grained extraction tasks, particularly sentence- or span-level classification of argumentative roles such as claims, premises, and counterarguments. Natural language processing (NLP) techniques are increasingly applied to legal documents for tasks such as judgment prediction, argument mining, and legal question answering [6]. Much of this work has focused on European Court of Human Rights (ECHR) cases [10] or Chinese legal corpora such as Chinese AI and Law Challenge (CAIL) [11]. Benchmarks such as LegalBench further explore higher-level reasoning tasks and evaluate the performance of LLMs in legal settings [7, 8], while recent work has also explored the development of domain-specific legal language models tailored to the U.S. legal system [12]. Although LLMs demonstrate strong generalization and flexibility, prior studies suggest that fine-tuned transformer models pretrained on legal corpora often remain more reliable and interpretable for classification tasks, especially under class imbalance and strict faithfulness requirements [13, 14]. Importantly, existing richly annotated LAM resources such as European Court of Human Rights - Argument Mining (ECHR-AM) and Legal Argument Mining: European Court of Human Rights (LAM: ECHR) focus primarily on non-U.S. jurisdictions,

***Corresponding author:** Haihua Chen, College of Information, University of North Texas, USA. haihua.chen@unt.edu

limiting their applicability to U.S. legal reasoning. Existing U.S.-focused legal NLP benchmarks largely emphasize downstream prediction or reasoning tasks rather than fine-grained argumentative role annotation. As a result, there remains no large-scale, systematically annotated corpus for sentence-level LAM across both federal and state U.S. caselaw.

To address these gaps, we introduce LAMUS, a large-scale LAM corpus constructed from U.S. Supreme Court decisions and Texas criminal appellate opinions. The dataset contains judicial opinions from 1921 to 2025 and approximately 2,900,083 sentences, each annotated with one of six argumentative roles: Fact, Issue, Rule/Law/Holding, Analysis, Conclusion, or Other. Building on prior work in semi-automated corpus construction [1], we adopt a data-centric approach that integrates corpus acquisition, automatic annotation using LLMs, targeted quality control, and systematic model evaluation.

Our objective is twofold: (1) to develop a scalable annotation pipeline that combines LLM-based automatic labeling with targeted human verification to improve label consistency and reliability and (2) to systematically evaluate how model scale, domain specialization, prompting strategy, and fine-tuning affect sentence-level legal argument classification performance.

This study is guided by the following research questions:

- RQ1: How effective are general-domain and legal-domain LLMs in performing sentence-level legal argument classification on U.S. caselaw?
- RQ2: How can different LLM prompting strategies (zero-shot, few-shot, and chain-of-thought [CoT]) be optimized to improve classification performance?
- RQ3: How can a scalable, high-quality LAM corpus be constructed from U.S. Supreme Court and state-level caselaw using semi-automated annotation?

To answer these questions, we construct the LAMUS corpus through large-scale case collection, preprocessing, sentence segmentation, LLM-based annotation, and targeted human verification. We then benchmark general-purpose and legal-domain language models under zero-shot, few-shot, and CoT prompting strategies, followed by fine-tuning experiments. Performance is evaluated using accuracy, precision, recall, and F1-score.

The main contributions of this work are as follows: (1) we present LAMUS, a large-scale sentence-level LAM corpus for U.S. Supreme Court and Texas criminal caselaw; (2) we introduce a scalable annotation pipeline combining LLM-based labeling with targeted human verification; and (3) we provide a systematic evaluation of prompting strategies, domain specialization, and fine-tuning for legal argument classification.

2. Related Work

2.1. Legal argument mining

LAM focuses on identifying and structuring argumentative components in judicial texts to support retrieval, summarization, and legal reasoning [3, 5, 15]. Earlier work relied on feature-based methods, while recent studies increasingly employ transformer-based architectures and sentence-level annotation schemes [16, 17]. Existing high-quality corpora, including ECHR-AM and LAM: ECHR, primarily focus on non-U.S. jurisdictions, while Chinese legal datasets such as CAIL target different legal systems and annotation objectives. As a result, large-scale sentence-level argument mining resources for U.S. caselaw remain limited.

Recent legal NLP research has also highlighted challenges including inconsistent annotation standards, class imbalance, and difficulty modeling long-range legal reasoning structures [5, 16]. These limitations motivate the development of scalable, high-quality U.S. LAM corpora with standardized annotation and verification procedures.

2.2. LLMs-as-Judge for domain applications

LLMs are increasingly used as evaluators in specialized domains, a paradigm commonly referred to as "LLMs-as-Judges" [18]. Prior work shows that LLMs can support scalable evaluation and quality control through natural language reasoning and classification. In legal NLP, recent studies have explored the use of LLM-based evaluation for legal reasoning and argument assessment, while also highlighting challenges related to bias, consistency, and prompt sensitivity [18–21]. These findings suggest that reliable deployment requires careful prompt design, domain adaptation, and human verification.

In this study, we apply the LLM-as-Judge paradigm as part of a semi-automated annotation and verification pipeline for sentence-level legal argument classification.

2.3. Legal intelligence with LLMs

LLMs are increasingly used in legal applications such as summarization, legal question answering, statute retrieval, judgment prediction, and document drafting [17]. Their ability to perform zero-shot and few-shot learning makes them attractive for legal tasks with limited labeled data, including LAM. Recent work demonstrates that combining LLMs with retrieval systems or structured legal knowledge can improve accuracy and reduce hallucinations.

At the same time, studies have raised concerns about the reliability of LLM reasoning in law. Extended or explicit reasoning does not always improve performance, particularly for smaller or non-specialized models [20].

Although LLMs enable scalable annotation and strong zero-shot performance, prior work shows that their outputs remain sensitive to prompting strategy and domain adaptation [16, 20]. Hybrid approaches combining LLM-based annotation with domain-specific models and human verification therefore remain important for reliable legal NLP systems. This study follows this direction through systematic evaluation of prompting strategies, fine-tuning, and targeted annotation refinement.

3. Research Methodology

3.1. Task definitions

LAM is the process of automatically identifying, classifying, and structuring argumentative components within legal texts. In this paper, our specific task is to extract legal arguments and categorize them into argument types, with a primary focus on argument classification.

1) Argument Component Detection – identifying spans of text that represent argumentative elements such as facts, issues, rules, analysis, and conclusions.
2) Argument Classification – assigning each detected span to a predefined category of argument type, forming the foundation for downstream reasoning tasks.

We formulate this task as a sentence-level classification problem. Each sentence in a judicial opinion is treated as an individual instance, and the model is trained to assign it to one of the following categories: Fact, Issue, Rule/Law/Holding, Analysis, Conclusion, or Other.

Table 1 summarizes the sentence-level annotation schema, including category definitions, representative examples, and annotation guidelines.

3.2. Framework

Figure 1 illustrates the end-to-end workflow adopted in this study for constructing the LAMUS corpus and evaluating LLMs for legal argument classification. The framework is designed as a modular, data-centric pipeline that integrates corpus acquisition, automatic annotation, quality control, and systematic model evaluation.

The workflow begins with large-scale corpus construction. Judicial opinions are collected from two primary sources: U.S. Supreme Court decisions obtained from Justia and state-level criminal caselaw from Texas. Raw case texts are cleaned through removal of metadata, Optical Character Recognition (OCR) artifacts, citation-only fragments, duplicated headers, and malformed text and then normalized and segmented into sentences using legal-domain preprocessing tools. Invalid or noisy segments, such as section headers, citation fragments, or malformed sentences, are filtered to ensure that each unit represents a meaningful argumentative span suitable for sentence-level classification.

Next, the cleaned sentences are passed to the LLM-based automatic annotation stage. Cleaned sentences are automatically annotated using standardized prompts derived from the annotation schema in Table 1. Models assign one of six categories: Fact, Issue, Rule/Law/Holding, Analysis, Conclusion, or Other. We evaluate zero-shot, few-shot, and CoT prompting strategies to examine how prompting affects classification performance.

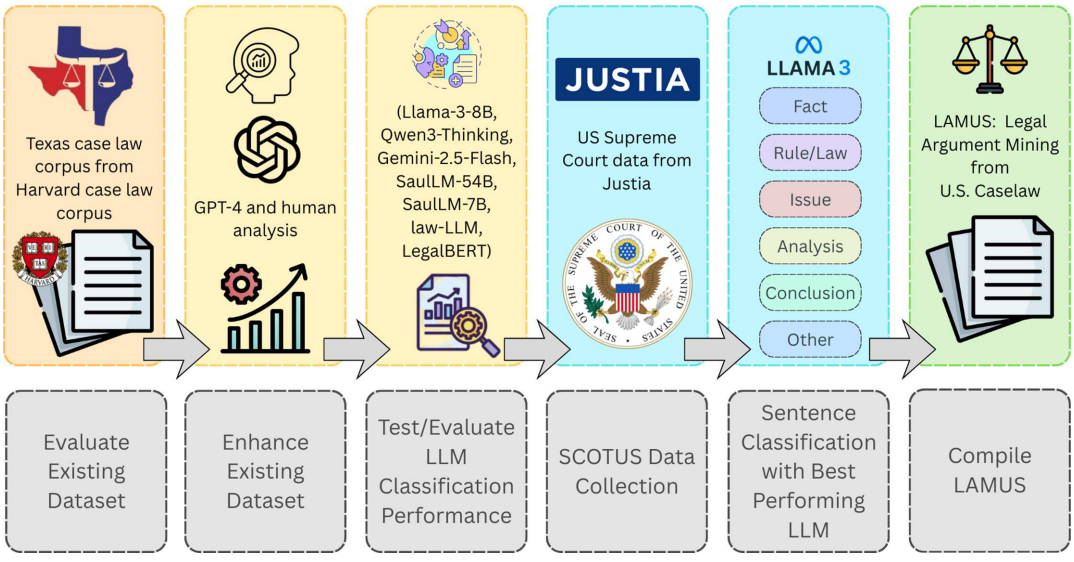
To improve annotation quality, model predictions were compared against existing human annotations for the Texas subset. Sentences with label disagreement or high-confidence alternative

Table 1
Annotation scheme for legal argument

Label	Description
Fact	Any fact that is pertinent to the case This includes testimony, statements of record, case history, and anything else that is a fact that helps establish the foundation of the case for the court to build its analysis and judgment on
Issue	Any issue or question that the court must decide. This includes the overall issue of the case as well as any sub-issues that are raised in the case
Rule/Law/Holding	Any statement of or reference to a rule, law, or holding. This includes sentences referencing a rule, law, or holding and then using it to reason through some point
Analysis	Any sentence that synthesizes information to further the court's reasoning. This includes sentences that refer to the facts of the case and then use them to push forward toward a resolution
Conclusion/Opinion/Answer	Any sentence that effectively resolves an issue facing the court
Others	(1) Any sentence or phrase that does not fit the other labels in terms of its content. This includes section headings and others. (2) Sentences that were not split correctly

Note: Please refer to our previous paper for examples and annotation guidelines for each category.

Figure 1
Methodology for corpus construction, LLM-based sentence annotation, data quality assessment, and model evaluation in the LAMUS dataset



predictions were flagged for manual review and corrected when necessary. This targeted verification strategy improves label consistency while preserving scalability.

Following data refinement, the enhanced dataset is used for model evaluation and benchmarking. Multiple LLMs with varying sizes and degrees of legal specialization are evaluated on the sentence-level classification task using standard metrics such as accuracy, precision, recall, and F1-score, reflecting growing efforts to benchmark LLM performance across legal reasoning and classification tasks [16, 22, 23] as well as the development of domain-adapted legal language models [24].

In addition to aggregate performance, label-level behavior and distributional patterns are analyzed to assess whether models preserve the logical flow of judicial reasoning.

The final output of this pipeline is the LAMUS corpus, a structured dataset of sentence-level legal arguments with verified annotations suitable for training and evaluating legal NLP models.

3.3. Data quality evaluation and improvement

Because the Texas criminal caselaw dataset was annotated by multiple human annotators, some degree of label inconsistency was unavoidable. Assigning labels such as Fact, Issue, Rule/Law/Holding, Analysis, and Other is inherently challenging in legal texts, as individual sentences may plausibly belong to more than one category. Differences in annotator interpretation therefore introduce noise and raise concerns about annotation quality. High-quality data is essential for training reliable machine learning models, as annotation errors can propagate through model predictions and degrade downstream performance [25, 26].

To address this issue, we incorporated a Generative Pre-trained Transformer (GPT)-based verification step to identify potentially mislabeled sentences [27]. The prompt template used for LLM-assisted verification of sentence annotations is shown in Figure 2. Sentences were flagged for review when the model's predicted label disagreed with the original annotation or when the model's confidence in an alternative label exceeded a predefined threshold. For each flagged instance, we manually re-examined the sentence in its full legal context and reassigned the label when appropriate. This hybrid approach—combining automated pre-screening with targeted human review—enabled efficient quality control without the need to re-annotate the dataset from scratch. Although residual annotation noise may remain because verification also relies partly on LLM judgments, manual review of flagged instances helps reduce systematic labeling errors while preserving scalability.

As a result, annotation consistency was substantially improved, with manual review correcting nearly one-fifth of flagged instances. This refinement enhances the reliability of downstream training and evaluation for legal argument classification, reduces annotation noise in subsequent experiments, and provides a stronger foundation for scaling argument mining methods to larger U.S. caselaw collections [25].

GPT-4 was used with the following prompt to flag potentially misclassified data.

3.4. Large language models

We evaluate seven LLMs that vary in scale, architecture, and degree of legal-domain specialization, as seen in Table 2, to assess their effectiveness for sentence-level legal argument classification [28].

Figure 2
Prompt template used for LLM-assisted verification of legal sentence annotations

Prompt for verifying legal annotations

System Prompt
You are a legal expert. Your task is to verify sentence annotations in U.S. legal cases. Return your response in JSON format with the following structure:

Sentence:	*<The sentence from the case>*
Assigned Label:	*<The label currently assigned>*
Correct?:	*<YES or NO>*
Explanation:	*<If NO, brief explanation and the correct label>*

Annotation Categories:

Fact	Pertinent events or records establishing the case's foundation; no reasoning.
Issue	Questions the court must decide; legal points derived from facts.
Rule/Law/Holding	References to laws, rules, or prior case holdings used to support reasoning.
Analysis	Reasoning synthesizing facts and law to progress toward a decision.
Conclusion	Resolves issues or states the court's final decision.
Others	Non-informative text or headings not fitting above categories.

User Prompt
Here is the sentence to verify: *<Sentence from the case>* with the assigned label: *<Assigned Label>*

Table 2
Large language models used in experiment

Model	Type	Params	Domain
LLaMA-3-8B	Decoder-only	8B	General
Qwen3-Thinking	Decoder-only	7B	General
Gemini-2.5-Flash	Application Programming Interface (API)-based	–	General
SaulLM-54B	Legal fine-tuned	54B	Legal
SaulLM-7B	Legal fine-tuned	7B	Legal
Law-LLM	Legal fine-tuned	7B	Legal
LegalBERT	BERT Encoder	110M	Legal

3.4.1. General-domain LLMs

General-domain LLMs are trained on broad, heterogeneous corpora spanning multiple domains, enabling strong general-purpose language understanding and reasoning, but without explicit domain specialization [22]. Prior work shows that such models can transfer effectively to specialized tasks through prompting, particularly when tasks emphasize general reasoning rather than domain-specific terminology [9, 29]. To evaluate their suitability for legal argument classification, we experiment with the following general-domain LLMs.

LLaMA-3-8B is an open-weight model developed by Meta that balances efficiency and reasoning capability, making it suitable for large-scale preprocessing and exploratory annotation tasks[1]. In this study, it is used to assess the feasibility of smaller open models for sentence-level legal argument classification.

Qwen3-Thinking is a reasoning-oriented LLM designed to support structured, multistep inference, which is particularly relevant for tasks requiring contextual interpretation and logical progression, such as LAM[2].

Gemini-2.5-Flash is a low-latency, high-throughput LLM developed by Google, optimized for fast inference and instruction following[3]. Although not legally specialized, its efficiency makes it suitable for large-scale automatic annotation pipelines.

3.4.2. Legal-domain LLMs

Domain-specific LLMs are trained or fine-tuned on field-specific corpora, enabling them to capture specialized vocabulary, discourse structures, and reasoning patterns that are underrepresented in general-domain data [22]. In the legal domain, exposure to statutes and caselaw has been shown to improve performance on legal NLP tasks requiring fine-grained argumentative distinctions [13]. The legal-domain models evaluated in this study are described below.

SaulLM-54B is a large-scale legal language model trained on extensive legal corpora, including statutes and judicial opinions, and is designed to capture complex legal reasoning patterns [30]. It serves as a high-capacity domain-adapted model for sentence-level legal argument classification.

Law-LLM represents a class of legal-domain LLMs optimized through pretraining or fine-tuning on legal texts [13]. In this study, it is evaluated for its ability to distinguish closely related argumentative categories.

SaulLM-7B is a smaller, computationally efficient variant of SaulLM that retains legal-domain specialization while reducing model size, allowing assessment of the trade-off between capacity and domain adaptation [30].

LegalBERT is a transformer-based model adapted from BERT and further pretrained on large-scale legal corpora to capture domain-specific language patterns [13]. It serves as a supervised transformer baseline for comparing prompted and fine-tuned LLM performance.

3.5. Prompt design

We evaluated three prompting strategies for legal sentence classification: few-shot prompting, zero-shot prompting, and CoT prompting. The complete prompt templates used in the experiments are shown in Figure 3(a)–(c).

The zero-shot prompt evaluates a model's intrinsic legal reasoning by providing only task instructions and label definitions, without any task-specific examples. The model relies entirely on pretraining to map legal sentences to categories such as Fact, Issue, Rule/Law/Holding, Analysis, Conclusion, or Other. Zero-shot prompting is widely used to assess generalization and latent task competence in LLMs [9, 31, 32], and in legal NLP, it helps determine whether a model has internalized abstract legal structures without supervision.

[1] Meta AI. 2024. The Llama 3 Model Family. https://ai.meta.com/llama/.

[2] Alibaba Cloud. 2025. Qwen3 Technical Report. https://qwenlm.github.io/. Reasoning-oriented large language models

[3] Google DeepMind. 2025. Gemini Model Family. https://ai.google.dev/gemini-api. Including Gemini 2.5 Flash.

Figure 3
Comparison of prompting strategies evaluated for legal sentence classification: (a) few-shot prompting with annotated examples, (b) zero-shot prompting using only task instructions and label definitions, and (c) chain-of-thought prompting incorporating explicit reasoning, justification, and confidence estimation

(a)

Few-Shot Legal Classification Prompt

System Prompt
You are a legal reasoning expert. Your task is to read a legal paragraph and classify each sentence into one of the following categories: Fact, Issue, Rule/Law/Holding, Analysis, Conclusion, Other. Follow this format:
Sentence: <text>
Label: <Fact | Issue | Rule/Law/Holding | Analysis | Conclusion | Other>
Annotation Categories:

Fact	Background details, case history, or objective evidence.
Issue	The legal question(s) the court must answer.
Rule/Law/Holding	Stated legal rules, statutes, or holdings from precedent.
Analysis	How the law applies to the facts; legal reasoning or interpretation.
Conclusion	The final decision or holding by the court.
Other	Anything that doesn't fall into the above categories.

User Prompt
Here is the paragraph to classify: *<Input Paragraph>*

(b)

Zero-Shot Legal Classification Prompt

System Prompt
You are a legal reasoning expert. Your task is to read a legal sentence and classify each sentence into one of the following categories: Fact, Issue, Rule/Law/Holding, Analysis, Conclusion, Other. Output a list of sentences with their assigned label.
Annotation Categories:

Fact	Objective background details, procedural history, or evidence.
Issue	The legal question(s) the court must resolve.
Rule/Law/Holding	Stated legal rules, statutes, or holdings from precedent.
Analysis	Application of law to fact; legal reasoning or interpretation.
Conclusion	Final decisions or determinations made by the court.
Other	Sentences that do not fit the above categories.

User Prompt
Here is the paragraph to classify: *<Input Paragraph>*

(c)

Chain-of-Thought Legal Classification Prompt

System Prompt
You are a legal reasoning expert. For each sentence in the paragraph:
(1) Assign exactly one label from: Fact, Issue, Rule/Law/Holding, Analysis, Conclusion, Other.
(2) Provide a short, explicit justification (1–3 numbered points) tied to textual cues.
(3) Provide a confidence score (0–100%) and a one-line reason for that confidence.

	Sentence:	<text>
	Label:	<one of the six>
Output Format:	Justification:	1. <explicit cue> 2. "<how it maps to label>"
	Confidence:	<%> — <explanation>

User Prompt
Here is the paragraph to classify: *<Input Paragraph>*

The few-shot prompt extends zero-shot prompting by including labeled examples for each legal argument category, providing explicit demonstrations of desired input–output behavior. This in-context learning approach allows the model to infer task structure and label semantics from a small number of exemplars [9, 23, 33]. In legal text classification, few-shot examples are especially useful because argument categories can be subtle and context-dependent, grounding abstract label definitions in concrete sentences to better align model predictions with established annotation schemes.

The structured CoT prompt builds on few-shot prompting by requiring explicit textual justifications and confidence scores for each classification. This encourages the model to articulate intermediate reasoning steps linking linguistic cues to the assigned label. CoT prompting has been shown to improve performance on complex reasoning tasks by decomposing decisions into interpretable steps [34–36]. In our setting, justifications are concise and text-grounded, enhancing transparency and auditability without revealing internal model deliberations.

3.6. Evaluation metrics

Model performance is evaluated using standard classification metrics commonly adopted in legal NLP. Accuracy measures the overall correctness of predictions across all classes. Precision quantifies the proportion of correctly predicted positive instances among all predicted positives. Recall (or sensitivity) measures the proportion of correctly predicted positive instances among all actual positives. F1-score is the harmonic mean of precision and recall.

4. Experiment and Results

4.1. Data collection

4.1.1. Initial dataset

The initial dataset for this study is derived from the Texas criminal caselaw corpus introduced by Chen et al. [1], constructed from the Harvard Law Library Case Law Corpus and previously used in our work [1]. This publicly available corpus spans approximately 360 years of U.S. judicial decisions across 582 reporters. From it, all published Texas Criminal Cases from 1840 to 2018 were extracted, yielding 27,712 cases stored in structured JSON format, with sentence-level data used as the unit of annotation.

Full texts were segmented into sentences using LexNLP, and invalid sentences—including headings, fragments, and segmentation errors—were filtered via a supervised binary classifier trained on manually labeled examples [37]. After preprocessing, 542,763 validated sentences were obtained, from which a subset of 5066 sentences was manually annotated as a carefully curated seed corpus for legal argument classification[4]. Annotation followed a six-class scheme grounded in the FIRAC framework: Fact, Issue, Rule/Law/Holding, Analysis, Conclusion/Opinion, and Other. Nine trained annotators (six undergraduate, three graduate) independently labeled each sentence using Doccano, with final labels determined by majority vote and adjudication when necessary. Inter-annotator agreement was measured using Cohen's kappa, and both high- and low-agreement annotations were retained, consistent with findings that moderate disagreement does not significantly impair downstream model performance [1].

This dataset was selected for two main reasons. First, its domain specificity ensures that argumentative structures reflect real-world judicial reasoning in U.S. criminal law, which often involves complex interactions between facts and precedent. Second, its high level of curation and quality validation makes it well suited for downstream applications such as argument classification and structure reconstruction. In this study, the Texas criminal caselaw dataset provides the foundation for refining annotation schemes, training and evaluating models, and benchmarking the effectiveness of LLMs in LAM.

4.1.2. Data quality improvement results/enhanced dataset

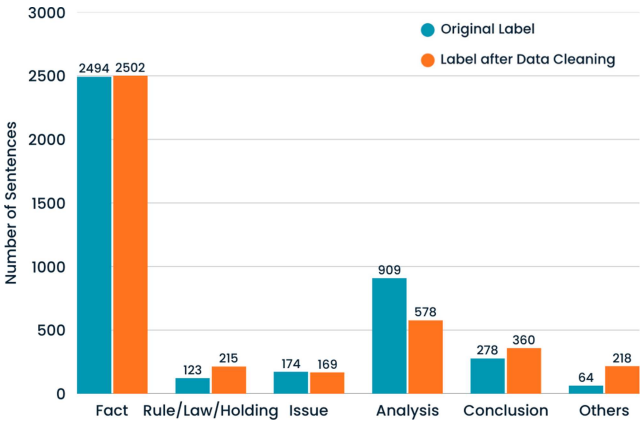
To assess and improve annotation quality, we conducted a systematic verification of the annotated dataset using a GPT-based label consistency check. Each of the 4042 annotated sentences was evaluated by comparing the original human-assigned label with GPT's predicted label. Sentences were flagged for further inspection when the model's predicted label differed from the original annotation or when the model expressed high confidence in an alternative classification. This automated screening step served to prioritize cases most likely to contain annotation errors.

Through this process, GPT flagged 1058 sentences as potentially mislabeled. All flagged sentences were then manually reviewed in their full legal context by human annotators. Of these cases, 273 sentences were confirmed to be incorrectly labeled and were subsequently corrected, while the remaining flagged sentences were determined to be correctly annotated. In addition to the model-flagged cases, targeted manual spot checks were conducted on unflagged portions of the dataset to estimate residual annotation noise.

Combining the confirmed corrections from the flagged set with errors identified during additional manual review, we estimate that approximately 785 sentences in total, corresponding to 19.4% of the dataset, carried incorrect labels prior to refinement. These findings highlight the prevalence of subtle annotation inconsistencies in sentence-level legal argument labeling and underscore the necessity of post-annotation quality control.

Following correction, the refined dataset exhibits substantially improved label consistency and reliability. Figure 4 illustrates the distributional comparison between the original dataset labels and the enhanced dataset labels after quality improvement, highlighting how the refinement process altered class proportions and reduced systematic mislabeling. The resulting enhanced dataset provides a reliable benchmark for training and evaluating legal argument classification models before applying them to large-scale corpus construction.

Figure 4
Comparison of Texas criminal caselaw dataset sentence labels before and after the data cleaning process, indicating improved quality



[4] https://github.com/haihua0913/legalArgumentmining

The final enhanced Texas dataset consists of 4042 sentence-level annotations spanning six legal argument categories. After quality refinement, approximately 19.4% of labels were corrected or reevaluated, substantially reducing annotation noise and improving inter-label consistency.

Importantly, the distribution of argument categories shifted following correction, indicating that certain classes—particularly Rule/Law/Holding and Analysis—were more prone to misclassification in the original annotations. This refined distribution better reflects the structural composition of judicial reasoning and provides a more reliable foundation for downstream model training and benchmarking.

4.1.3. Final LAMUS corpus

The final LAMUS corpus consists exclusively of U.S. Supreme Court (SCOTUS) opinions automatically labeled for sentence-level legal argument structure. The corpus was constructed using an LLM (LLaMA-3-70B) that was selected based on its strong performance in the preceding classification experiments.

The refined Texas criminal case dataset was used as a high-quality benchmark and training resource for evaluating legal argument classification models. After identifying the most effective model configuration, we applied the selected model to automatically label sentences from the SCOTUS opinion corpus.

Following sentence segmentation and preprocessing, the resulting LAMUS corpus contains 2,900,083 labeled sentences spanning Supreme Court decisions from 1921 through 2025. Each sentence is annotated with one of six argument categories: Fact, Issue, Rule/Law/Holding, Analysis, Conclusion, and Other.

This large-scale corpus provides a comprehensive resource for studying judicial reasoning and supports downstream tasks such as LAM, argument reconstruction, and computational analysis of Supreme Court opinions.

4.2. Experiment design and settings

To evaluate the effectiveness of LLMs for sentence-level legal argument classification, we designed a series of experiments that systematically varied model prompting strategies, fine-tuning, and model scale.

LLM with Different Prompting Strategies: We tested each model under three distinct prompting conditions: zero-shot, few-shot, and CoT. In the zero-shot setting, models were asked to label sentences without any examples. Few-shot prompting provided labeled sentences ranging from 1 to 100 examples, systematically testing 1, 3, 4, 5, 10, 20, 40, 60, 80, and 100 examples to determine the optimal count. Examples were drawn from a curated set of 100 legal excerpts organized into 20 groups of 5 sentences (one per category). CoT prompting required models to reason step by step, encouraging systematic feature identification before producing a label. This setup allowed us to investigate how prompting interacts with model scale and domain specialization.

LLM Fine-Tuning: For certain legal-domain models, we experimented with fine-tuning on 2585 labeled sentences from our corpus using QLoRA (4-bit quantization) with systematic hyperparameter optimization across learning rates, LoRA ranks, and training epochs. This allowed us to examine whether task-specific adaptation could improve sentence-level classification accuracy beyond what is achievable with prompting alone.

LLM with Different Parameter Scales: To explore the impact of model capacity, we included LLMs ranging from 2.5 billion to 54 billion parameters. By comparing performance across scales, we assessed the extent to which larger models benefit from multistep reasoning (CoT) and how smaller models may be limited in representational depth when performing complex classification tasks.

4.3. Evaluation results

4.3.1. Overall results

Table 3 summarizes classification accuracy across models and prompting strategies. CoT prompting achieved the strongest prompted performance, with LLaMA-3-8B reaching 75.89% accuracy. SaulLM-54B improved from 67.39% under zero-shot prompting to 72.80% with CoT prompting, while smaller legal-domain models such as SaulLM-7B performed better under zero-shot settings. Few-shot prompting consistently underperformed and frequently reduced accuracy relative to zero-shot baselines. Gemini-2.5-Flash performed near-randomly across all prompting conditions. Figure 5 visualizes the experimental results across all models and settings.

Overall, these results demonstrate that large-scale models combined with CoT prompting can substantially improve sentence-level legal argument classification, emphasizing the importance of aligning prompting strategy with both model capacity and domain specialization.

4.3.2. Performance of different prompting strategies

Among the evaluated models, the effectiveness of prompting strategies varies significantly. CoT prompting outperformed zero-shot and few-shot approaches for large, general-purpose models, such as LLaMA-3-8B, indicating that explicit stepwise reasoning helps models systematically identify relevant sentence features. Smaller models, including SaulLM-7B, were better suited to zero-shot prompting due to limited reasoning capacity.

Table 3
Zero-shot, few-shot, and chain-of-thought (CoT) performance of evaluated models. Bold indicates the best performance. Stability was assessed across 10 independent runs for the best-performing prompted configuration (mean 74.71% ± 0.56%)

Model	Domain	Zero-shot (%)	Few-shot (%)	CoT (%)
LLaMA-3-8B	General	65.38	45.75	75.89
SaulLM-54B	Legal	67.39	64.76	72.80
Law-LLM	Legal	60.12	31.68	28.75
Qwen3-Thinking	General	56.11	49.30	54.10
SaulLM-7B	Legal	52.09	21.64	38.02
Gemini-2.5-Flash	General	5.41	5.41	5.41

Figure 5
Overall comparison of experimental results across all models and settings, including baseline performance, accuracy percentages, and indicators for fine-tuned configurations

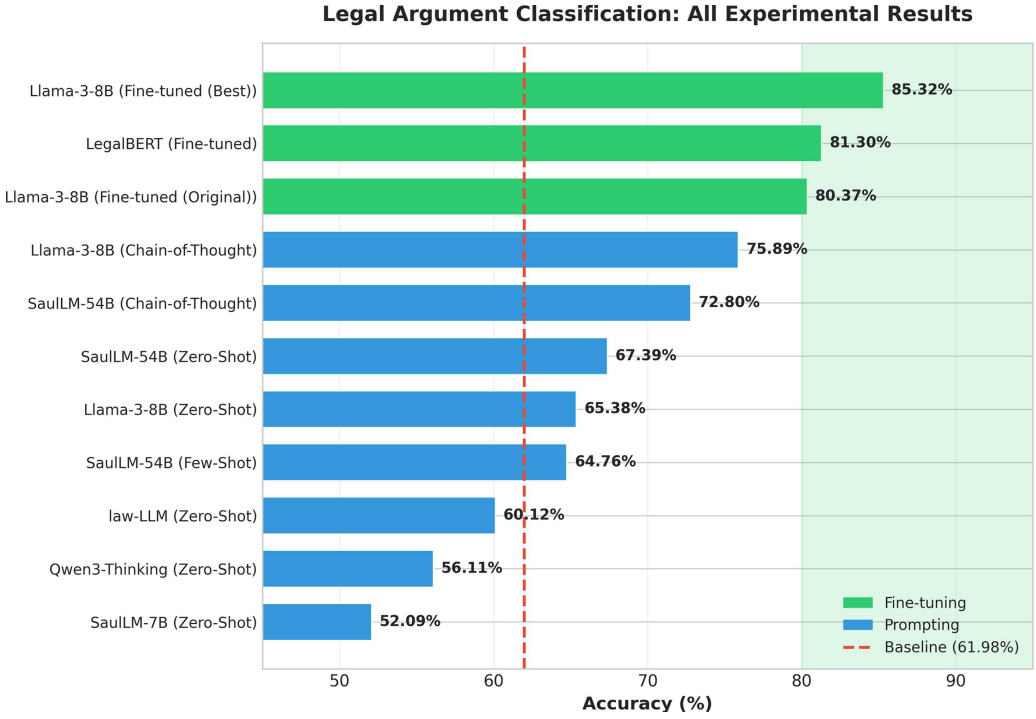


Table 4
Impact of few-shot prompting on model performance across completed experiments

Model	Zero-shot (%)	1-ex (%)	3-ex (%)	4-ex (%)	5-ex (%)
LLaMA-3-8B	65.38	47.76	49.61	52.86	50.54
SaulLM-54B	67.39	54.40	66.31	59.51	67.70

These observations highlight that CoT is not universally beneficial and interacts strongly with both model scale and domain expertise.

4.3.3. The impact of few-shot samples

Few-shot prompting consistently reduced performance for the models fully evaluated in this study. Table 15 reports results for LLaMA-3-8B and SaulLM54B, the two models for which all prompting configurations were completed.

For LLaMA-3-8B, few-shot prompting substantially degraded accuracy relative to the zero-shot baseline, with decreases ranging from 12% to 17% depending on the number of examples provided. Increasing the number of examples did not recover zero-shot performance, suggesting that few-shot demonstrations may introduce noise or encourage overfitting to prompt structure rather than improve task understanding. SaulLM-54B exhibited more stable behavior, with accuracy remaining broadly comparable to the zero-shot baseline across few-shot settings. This contrast suggests that the negative impact of few-shot prompting is more pronounced for general-purpose models than for larger legal-domain models.

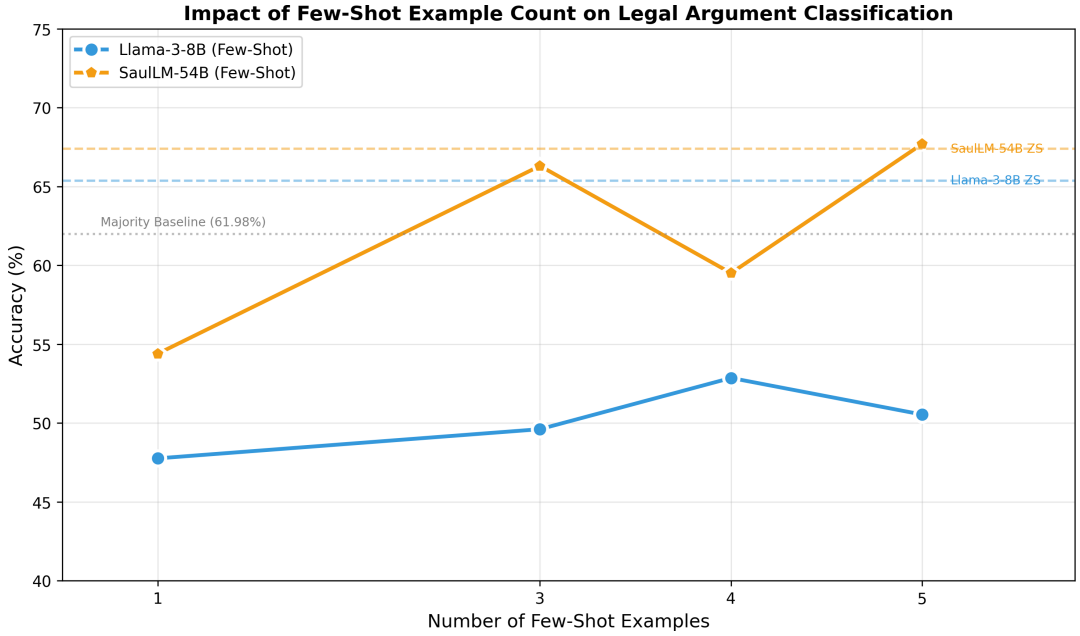
To further investigate the relationship between example count and classification accuracy, we extended the analysis using LLaMA-3-8B with 5, 10, 20, 40, 60, 80, and 100 examples (Table 5). As shown in Figure 6, performance degradation became more pronounced as the number of examples increased. Accuracy declined from 67.23% under zero-shot prompting to 65.07% with 5 examples and continued decreasing to 53.94% with 100 examples, representing a 13.29 percentage-point reduction from the baseline.

Table 5
Extended few-shot analysis (5–100 examples) for LLaMA-3-8B. Performance degrades monotonically as example count increases

# Examples	Accuracy (%)	Δ vs zero-shot
0 (zero-shot)	67.23	Baseline
5	65.07	–2.16%
10	66.15	–1.08%
20	64.91	–2.32%
40	65.53	–1.70%
60	60.43	–6.80%
80	59.04	–8.19%
100	53.94	–13.29%

*Zero-shot baseline varies slightly between experiments (65.38% vs 67.23%) due to differences in prompt formatting. Table 4 uses the original prompt template, while Table 5 uses a standardized template for fair comparison across the 5–100 example sweeps.

Figure 6
Accuracy of LLaMA-3-8B and SaulLM-54B under varying few-shot example counts. Few-shot prompting harms performance for LLaMA-3-8B, while SaulLM-54B remains relatively stable



These findings indicate that fixed few-shot demonstrations may interfere with generalization rather than improve legal reasoning performance in domain-specific classification tasks. One possible explanation is domain mismatch, where the curated examples do not adequately capture the jurisdiction-specific terminology and reasoning patterns present in Texas criminal appellate opinions.

4.3.4. Fine-tuning

Fine-tuning substantially outperformed all prompting-based approaches and produced the strongest overall classification performance. LLaMA-3-8B achieved the highest accuracy of 85.32% after fine-tuning, representing a 23.18% improvement over the majority-class baseline. LegalBERT also performed well at 81.30%, confirming that domain-specialized models benefit from task-specific adaptation. Even the original fine-tuned LLaMA-3-8B (without ablation) surpassed 80%, showing that careful hyperparameter tuning can yield further gains as seen in Table 6.

Table 6
Accuracy of fine-tuned models and improvement over baseline

Model	Method	Accuracy (%)	Baseline (%)
LLaMA-3-8B	Fine-tuned (ablation best)	85.32	+23.34
LegalBERT	Fine-tuned	81.30	+19.32
LLaMA-3-8B	Fine-tuned (original)	80.37	+18.39

Table 7 provides per-class performance for LegalBERT. The model achieves the highest precision and recall on the "Facts" category, while performance drops on "Rule/Law/Holding" and "Others," highlighting the challenges of classifying nuanced or infrequent legal arguments.

4.3.5. Ablation study

The ablation study (Figure 7 and Table 8) shows the effect of hyperparameter choices on LLaMA-3-8B. The best configuration (learning rate le-4, 5 epochs, LoRA rank 8) achieved 85.32% accuracy. Learning rate had the largest impact (±10%), while epochs and LoRA rank were moderately or minimally sensitive. Table 9 summarizes the sensitivity of each hyperparameter.

Table 7
Per-class performance metrics for LegalBERT on sentence-level legal argument classification

Category	Precision	Recall	F1	Support
Facts	0.91	0.94	0.92	401
Issue	0.86	0.70	0.78	27
Rule/Law/ Holding	0.79	0.32	0.46	34
Analysis	0.56	0.67	0.61	92
Conclusion	0.67	0.76	0.71	58
Others	0.70	0.40	0.51	35
Weighted Avg	0.82	0.81	0.81	647

4.4. Discussions

4.4.1. Why chain-of-thought prompting outperforms few-shot prompting

The observed performance gains from CoT prompting can be attributed to its ability to explicitly induce step-by-step reasoning within the model. By requiring intermediate reasoning steps, CoT encourages systematic identification of legally relevant features—such as factual predicates, normative rules, and analytical conclusions—rather than allowing the model to jump directly to a label. In contrast, few-shot prompting tends to promote

Figure 7
Ablation study: learning rate vs accuracy by LoRA rank and training epochs. Results from 36 experiments across 4 learning rates (1e-5, 5e-5, 1e-4, 2e-4), 3 LoRA ranks (8, 16, 32), and 3 epoch settings (1, 3, 5). The best configuration (LR = 1e-4, Rank = 8, Epochs = 5) achieves 85.32% accuracy

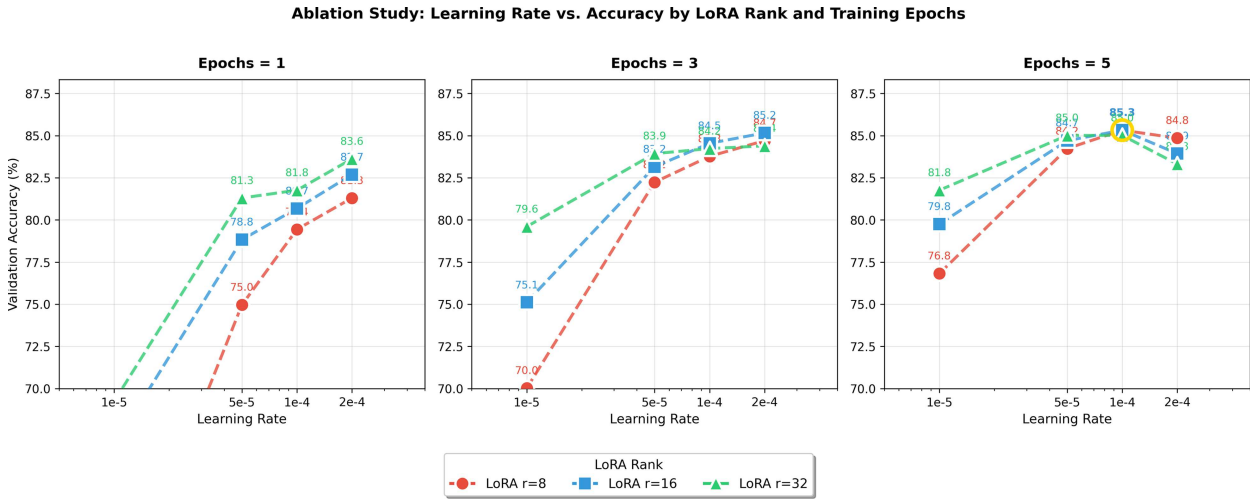


Table 8
Complete ablation study results for fine-tuned LLaMA-3-8B. The best configuration is highlighted

Config	Learning rate	Epochs	LoRA rank	Accuracy (%)
best	1e-4	5	8	85.32
lr_1e-4	1e-4	5	16	85.32
lr_2e-4	2e-4	3	16	85.16
lora_32	5e-5	5	32	85.01
epoch_5	1e-4	5	32	85.01
lora_8	2e-4	5	8	84.85
lr_2e-4	2e-4	3	8	84.70
lora_16	5e-5	5	16	84.70
lr_1e-4	1e-4	3	16	84.54
lora_32	2e-4	3	32	84.39

Table 9
Hyperparameter sensitivity analysis for fine-tuned LLaMA-3-8B based on 36 ablation experiments

Parameter	Sensitivity	Optimal value	Impact on accuracy (%)
Learning rate	HIGH	1e-4	±10
Epochs	MODERATE	5	±2.5
LoRA rank	LOW	8	±0.8

surface-level pattern matching, where the model aligns new inputs with previously seen examples without fully internalizing the underlying classification criteria. This distinction is particularly salient for sentence-level legal argument classification, where subtle semantic differences separate closely related categories.

Model capacity further mediates the effectiveness of CoT prompting. Our results indicate that only sufficiently capable models (≥8B parameters) consistently benefit from CoT, suggesting that smaller models lack the representational depth required to maintain coherent multistep reasoning. For these models, the added reasoning burden may introduce noise rather than clarity, diminishing any potential gains. This suggests that CoT effectiveness depends strongly on model scale and reasoning capacity.

4.4.2. Why general-purpose models outperform legal-domain models

Despite their domain specialization, several legal-domain language models underperform the strongest general-purpose models evaluated in this study such as LLaMA3-8B-Instruct in this task. One explanation lies in differences in training data quality and diversity. LLaMA models are trained on substantially larger and more heterogeneous corpora, enabling broader semantic coverage and more robust generalization. This diversity appears advantageous for sentence-level classification, which requires flexible interpretation of legal language across varying contexts rather than strict adherence to domain-specific templates.

Additionally, instruction-tuned general models demonstrate superior adherence to task instructions. LLaMA-3-8B-Instruct, in particular, is explicitly optimized for following structured prompts, which likely enhances its responsiveness to CoT and classification directives. In contrast, legal-domain models may be overfitted to other legal NLP objectives—such as document retrieval or citation prediction—leading to reduced adaptability when confronted with fine-grained argument role classification. This specialization may reduce generalization beyond pretraining distributions.

4.4.3. Why fine-tuning produces substantial performance gains

Fine-tuning yields the largest performance improvement (+9.27%), underscoring its effectiveness for this task. Through supervised fine-tuning, the model directly learns task-specific vocabulary, stylistic cues, and semantic patterns associated with each argumentative category. This process sharpens decision boundaries between conceptually similar labels, such as Rule and Analysis, which are particularly challenging to distinguish using prompting alone.

Moreover, fine-tuning resolves several practical limitations of prompt-based approaches. By eliminating reliance on free-form generation, it reduces output parsing errors and ensures consistent label formatting. The use of label masking further concentrates learning on the classification objective, preventing the model from expending capacity on irrelevant token prediction.

Together, these factors explain the pronounced gains achieved through fine-tuning and highlight its importance for high-precision legal argument classification. Although the results demonstrate strong performance improvements, residual annotation noise and jurisdiction-specific language patterns may still affect generalization to broader legal domains.

Despite these improvements, several limitations remain. First, because portions of the corpus are automatically labeled using LLMs, residual annotation errors and model-induced biases may persist despite human verification. Second, the training and evaluation data are primarily derived from Texas criminal appellate opinions and SCOTUS cases, which may limit generalization to other jurisdictions, practice areas, or legal systems. Finally, GPU availability constraints prevented complete evaluation of all prompting configurations across every model, particularly for extended few-shot experiments.

5. LAMUS Corpus Construction Using LLMs

The construction pipeline therefore proceeds in two stages: (1) model evaluation using the manually curated Texas criminal case dataset and (2) large-scale automatic labeling of SCOTUS opinions to construct the LAMUS corpus.

5.1. Data sources

For this study, we constructed the LAMUS corpus, a comprehensive dataset of U.S. Supreme Court (SCOTUS) opinions spanning from 1759 to the present. To compile the corpus, we developed a custom web scraper that systematically retrieved case texts from the Justia legal database, which provides open access to court opinions. The dataset includes all reported cases, beginning with early Pennsylvania cases and continuing through the Roberts Court, ensuring coverage of the full historical breadth of the U.S. Supreme Court.

Each case in the corpus contains the full opinion text, including majority, concurring, and dissenting opinions where available. Metadata such as the case name, citation, date of decision, and participating justices were also extracted to enable structured analysis.

The resulting corpus represents one of the most extensive publicly accessible compilations of Supreme Court decisions, providing a rich resource for research in LAM, NLP, and computational legal studies.

5.2. Data labeling using LLMs

The SCOTUS corpus was automatically annotated using LLaMA-3-70B, selected based on its strong classification performance in earlier experiments. Each sentence was assigned one of six argumentative roles: Fact, Issue, Rule/Law/Holding, Analysis, Conclusion, or Other. This approach enabled scalable sentence-level annotation across the full corpus while preserving consistency in legal reasoning structure.

5.3. Human verification

To validate the quality of our LLM-based automatic labeling pipeline, we conducted human verification following standard inter-annotator agreement protocols. Two expert annotators independently labeled a stratified random sample of 600 sentences from the SCOTUS labeled dataset, with 100 sentences per argument category ensuring balanced representation across all six labels.

Both annotators received identical instructions consistent with the annotation guidelines in Table 1 and completed labeling independently. Inter-annotator agreement was measured using Cohen's kappa (κ), which accounts for agreement occurring by chance. Cohen's kappa is computed as $\kappa = (p_o - p_e) / (1 - p_e)$, where p_o is the observed agreement between annotators and p_e is the expected agreement by chance.

Table 10 summarizes the human verification results. The annotators achieved a Cohen's kappa of $\kappa = 0.85$, indicating almost perfect agreement according to standard interpretation scales [38, 39]. Direct agreement was 87.3%, with only 76 sentences (12.7%) showing disagreement, primarily in edge cases involving sentences with characteristics of multiple categories.

Agreement between human annotators and model predictions averaged 89.2%, with Annotator 1 agreeing with model predictions in 90.5% of cases and Annotator 2 in 87.8%. These results validate the reliability of the automated labeling pipeline and confirm that the LAMUS corpus maintains annotation quality suitable for downstream legal NLP research.

Table 10
Human verification results for SCOTUS labeled data (n = 600 sentences)

Metric	Value
Total sentences verified	600
Sentences per category	100
Number of annotators	2
Cohen's kappa (κ)	0.85
Interpretation	Almost perfect
Direct agreement	87.3%
Annotator 1 vs model	90.5%
Annotator 2 vs model	87.8%
Average human-model	89.2%
Total disagreements	76 (12.7%)

5.4. Statistics of the corpus

The LAMUS corpus represents one of the largest structured resources for sentence-level LAM in U.S. caselaw. It combines a high-quality manually annotated dataset derived from Texas criminal appellate decisions with a large-scale automatically labeled corpus of U.S. Supreme Court opinions.

After preprocessing and sentence segmentation, the SCOTUS portion of the corpus contains 2,900,083 labeled sentences across six argument categories. These sentences span eight Supreme Court eras from the Taft Court (1921) through the Roberts Court (2025), providing substantial historical coverage for studying the evolution of legal reasoning.

Figures 8 and 9 illustrate the distribution of legal argument categories across the full corpus and across Supreme Court eras. Table 11 summarizes the distribution of labeled sentences by court era. The Burger Court (1969–1986) constitutes the largest portion of the dataset, followed by the Rehnquist and Warren Courts. More recent decisions from the Roberts Court (2005–2025) account for 362,891 labeled sentences.

The corpus captures substantial temporal diversity in legal reasoning and argumentative style across more than a century of jurisprudence. All labeled data are released publicly to support reproducibility and future research, including both a complete

Figure 8

Overall distribution of legal argument label categories across all Supreme Court eras (1921–present), comprising 2,900,083 sentences. The corpus is dominated by Analysis, Rule/Law/Holding, and Facts, while Issue and Conclusion occur less frequently but consistently

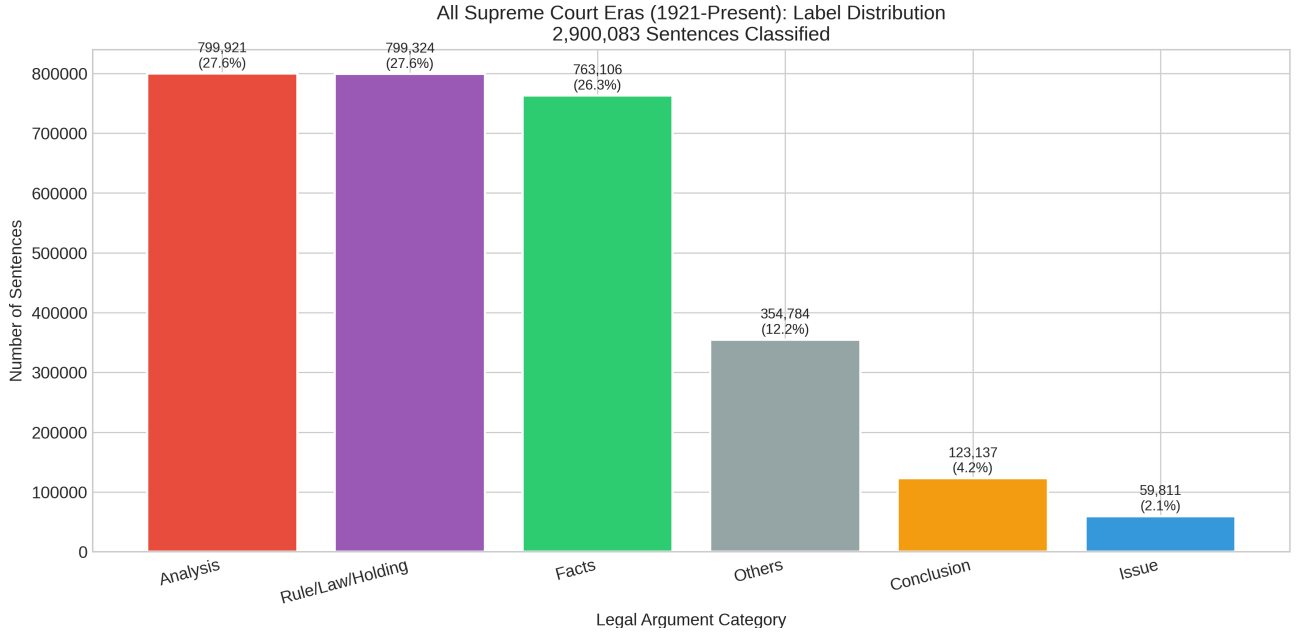


Figure 9

Distribution of legal argument label categories across Supreme Court eras. Each cell shows the percentage of sentences assigned to a label within a given court era

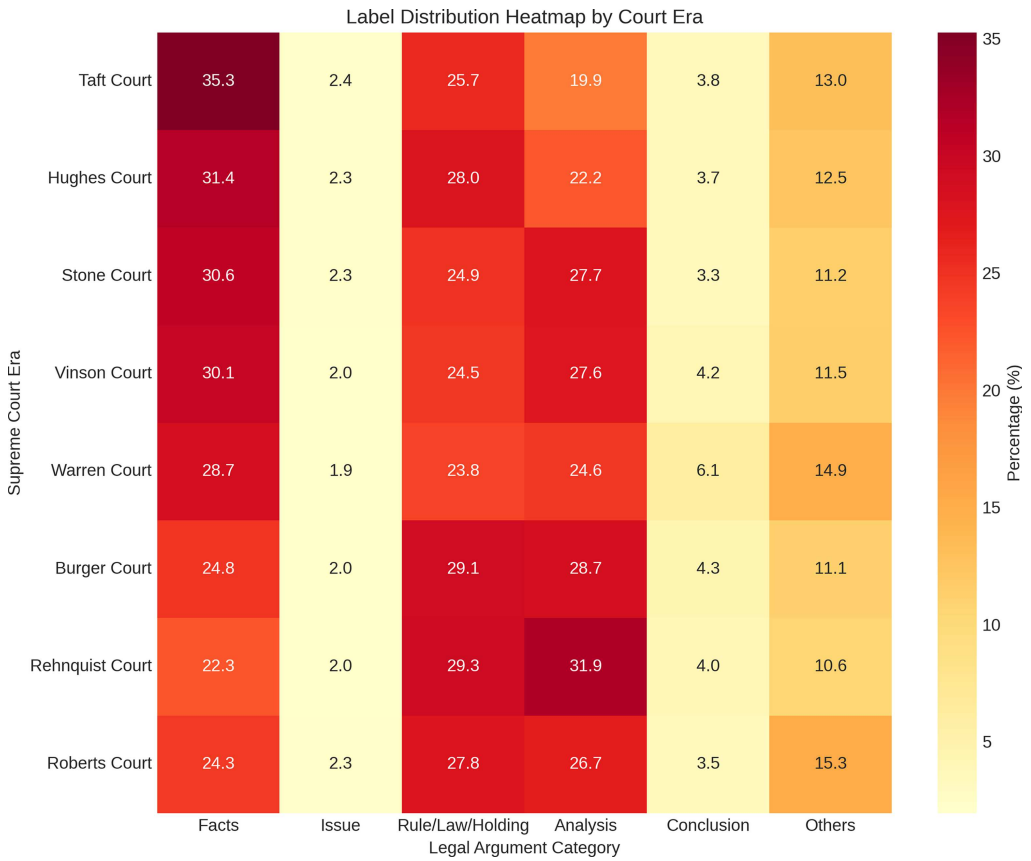


Table 11
Distribution of labeled sentences across Supreme Court eras in the LAMUS corpus

Court era	Years	Sentences
Burger Court	1969–1986	809,409
Rehnquist Court	1986–2005	673,564
Warren Court	1953–1969	377,645
Roberts Court	2005–2025	362,891
Hughes Court	1930–1941	213,122
Vinson Court	1946–1953	170,975
Taft Court	1921–1930	155,066
Stone Court	1941–1946	137,411
Total	1921–2025	2,900,083

corpus and a Roberts Court–only subset for focused analysis of modern jurisprudence.

Figure 10 illustrates a substantial domain shift between the Texas Criminal Cases training corpus and the SCOTUS dataset. The Texas corpus is heavily dominated by Facts (61.9%), whereas the SCOTUS dataset exhibits a more balanced distribution, with Rule/Law/Holding and Analysis each accounting for 27.6% of the data. This contrast highlights differing argumentative emphases between state-level criminal opinions and U.S. Supreme Court decisions.

5.5. Implications

The LAMUS corpus provides a foundational resource for computational legal research and the development of legal AI systems, particularly for tasks involving argument generation, argument reconstruction, case summarization, legal question answering, and predictive modeling of judicial decisions.

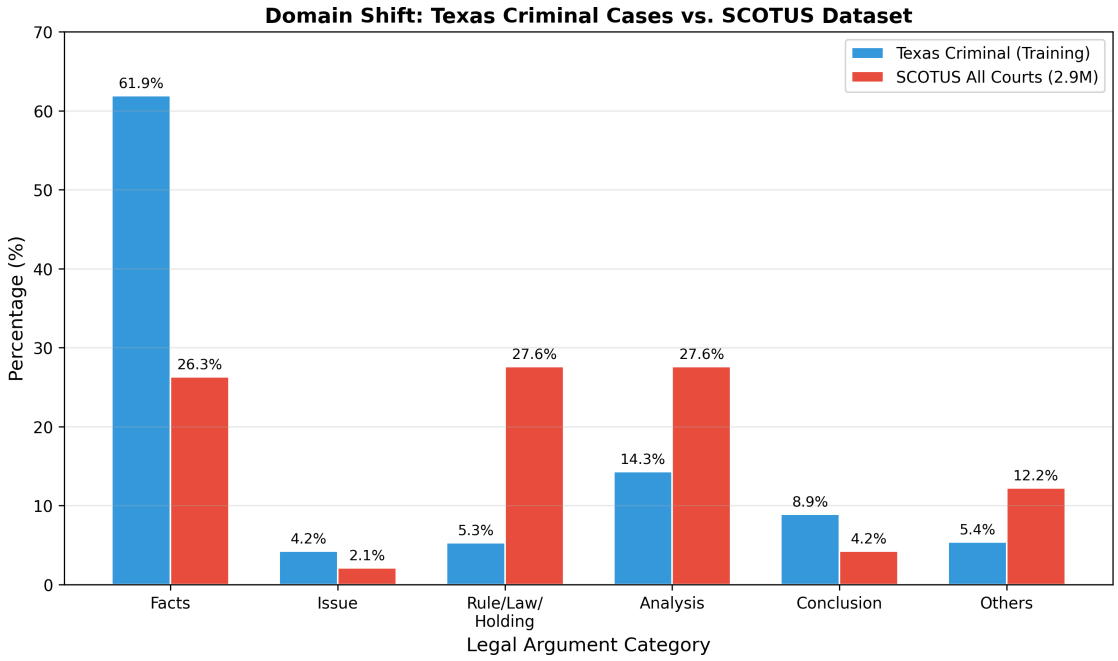
By combining full opinion texts with structured argumentative labels and metadata, the corpus enables analysis of legal reasoning patterns across Supreme Court eras and supports downstream applications in legal NLP. Its public availability also supports reproducibility and benchmarking in computational legal studies.

6. Conclusions

This study investigated the effectiveness of LLMs for sentence-level legal argument classification through systematic evaluation of prompting strategies, model scale, and supervised fine-tuning. Using a refined Texas criminal caselaw dataset, we compared zero-shot, few-shot, and CoT prompting approaches across multiple legal-domain and general-purpose models. The results show that prompting effectiveness depends strongly on both model capacity and domain specialization. CoT prompting improved performance for sufficiently large models, while few-shot prompting consistently reduced accuracy across completed experiments. Fine-tuning produced the strongest overall results, with fine-tuned LLaMA-3-8B achieving 85.32% accuracy, outperforming all prompt-based approaches. Building on these findings, we constructed the LAMUS corpus, containing 2,900,083 sentence-level annotations from U.S. Supreme Court opinions spanning 1921–2025 and validated the reliability of the automated labeling pipeline through human verification achieving Cohen's kappa of 0.85.

The findings of this work have important implications for computational legal research and legal AI development. First, the results demonstrate that model adaptation through fine-tuning is substantially more effective than prompt engineering alone for high-accuracy legal argument classification, particularly under domain shift between training and deployment corpora. Second, the consistent degradation observed with few-shot prompting suggests that fixed demonstration examples may introduce noise in specialized legal classification tasks rather than improve reasoning performance. Third, the construction of the LAMUS corpus provides one of the largest publicly available resources for sentence-level LAM, enabling downstream applications such

Figure 10
Comparison of legal argument label distributions in the Texas Criminal Cases training corpus and the SCOTUS dataset, highlighting substantial domain shift in argumentative structure across courts



as legal argument reconstruction, case summarization, precedent analysis, and legal question answering. More broadly, this work highlights the importance of annotation quality, human verification, and careful evaluation when deploying LLM-based systems in high-stakes legal settings where reliability and interpretability are essential.

Future work will focus on improving both model robustness and corpus quality. Additional experiments should evaluate broader combinations of prompting strategies and larger collections of legal-domain and general-purpose models, including adaptive retrieval-based few-shot prompting and jurisdiction-specific example selection to better address domain mismatch. Further refinement of the annotation process through increased involvement of legal experts may reduce residual annotation noise and improve dataset consistency. Beyond classification accuracy, future research should investigate practical applications of structured legal argument representations for automated argument generation, legal discourse analysis, precedent retrieval, and decision-support systems. Because legal AI systems may influence high-stakes decisions, future deployment should also incorporate transparency mechanisms, human oversight, and careful analysis of potential biases across jurisdictions, court levels, and case types.

Funding Support

This work was supported in part by The United States National Science Foundation (NSF) under Grant #2225229.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

Haihua Chen is the Editorial Board Member for the *Journal of Computational Law and Legal Technology* and was not involved in the editorial review or the decision to publish this article. The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in LAMUS at https://github.com/LavanyaPobbathi/LAMUS/tree/main.

Author Contribution

Serene Wang: Methodology, Validation, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Lavanya Pobbathi:** Software, Validation, Formal analysis, Investigation, Data curation, Writing – review & editing, Visualization. **Haihua Chen:** Conceptualization, Methodology, Validation, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

References

[1] Chen, H., Pieptea, L. F., & Ding, J. (2022). Construction and evaluation of a high-quality corpus for legal intelligence using semiautomated approaches. *IEEE Transactions on Reliability*, *71*(2), 657–673. https://doi.org/10.1109/TR.2022.3156126

[2] Mochales, R., & Moens, M.-F. (2011). Argumentation mining. *Artificial Intelligence and Law*, *19*(1), 1–22. https://doi.org/10.1007/s10506-010-9104-x

[3] Mochales Palau, R., & Moens, M.-F. (2009). Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, 98–107. https://doi.org/10.1145/1568234.1568246

[4] Zhang, G., Nulty, P., & Lillis, D. (2022). Enhancing legal argument mining with domain pre-training and neural networks. *Journal of Data Mining & Digital Humanities*. https://doi.org/10.46298/jdmdh.9147

[5] Ariai, F., Mackenzie, J., & Demartini, G. (2025). Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges. *ACM Computing Surveys*, *58*(6), 1–37. https://doi.org/10.1145/3777009

[6] Shu, D., Zhao, H., Liu, X., Demeter, D., Du, M., & Zhang, Y. (2024). LawLLM: Law large language model for the U.S. legal system. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 4882–4889. https://doi.org/10.1145/3627673.3680020

[7] Guha, N., Nyarko, J., Ho, D., Ré, C., Chilton, A., Chohlas-Wood, A., . . . , & Li, Z. (2023). 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*, 44123–44279. https://dl.acm.org/doi/10.5555/3666122.3668037

[8] Li, H., Chen, J., Yang, J., Ai, Q., Jia, W., Liu, Y., . . . , & Huang, M. (2025). LegalAgentBench: Evaluating LLM agents in the legal domain. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, (*Volume 1: Long Papers*), 2322–2344. https://doi.org/10.18653/v1/2025.acl-long.116

[9] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., . . . , & Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 27730–27744. https://dl.acm.org/doi/10.5555/3600270.3602281

[10] Poudyal, P., Šavelka, J., Ieven, A., Moens, M.-F., Goncalves, T., & Quaresma, P. (2020). ECHR: Legal corpus for argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, 67–75.

[11] Zhang, H., & Dou, Z. (2023). Case retrieval for legal judgment prediction in legal artificial intelligence. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, 801–812.

[12] Shen, J., Xu, J., Hu, H., Lin, L., Ma, G., Zheng, F., . . . , & Han, W. (2025). A law reasoning benchmark for LLMs with tree-organized structures including factum probandum, evidence, and experiences. In *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 17252–17274). https://doi.org/10.18653/v1/2025.findings-acl.887

[13] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The muppets straight out of law school. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 2898–2904). https://doi.org/10.18653/v1/2020.findings-emnlp.261

[14] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., . . . , & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1–38. https://doi.org/10.1145/3571730

[15] Nguyen, H. T., Fungwacharakorn, W., Zin, M. M., Goebel, R., Toni, F., Stathis, K., & Satoh, K. (2025). LLMs for legal reasoning: A unified framework and future perspectives. *Computer Law & Security Review*, *58*, 106165. https://doi.org/10.1016/j.clsr.2025.106165

[16] Chi, X., Wang, W., Zhang, Z., Li, A., Huang, Y., Wu, Y., . . . , & Xiong, M. (2026). LegalAI research in LLM era: data, modeling and evaluation. *Artificial Intelligence Review*, *59*, 118. https://doi.org/10.1007/s10462-026-11514-9

[17] Siino, M., Falco, M., Croce, D., & Rosso, P. (2025). Exploring LLM applications in law: A literature review on current legal NLP approaches. *IEEE Access*, *13*, 18253–18276. https://doi.org/10.1109/ACCESS.2025.3533217

[18] Li, D., Jiang, B., Huang, L., Beigi, A., Zhao, C., Tan, Z., . . . , & Liu, H. (2025). From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2757–2791. https://doi.org/10.18653/v1/2025.emnlp-main.138

[19] Enguehard, J., Van Ermengem, M., Atkinson, K., Cha, S., Chowdhury, A. G., Ramaswamy, P. K., . . . , & Mincu, D. (2025). LeMAJ (Legal LLM-as-a-judge): Bridging legal reasoning and LLM evaluation. In *Proceedings of the Natural Legal Language Processing Workshop 2025*, 318–337. https://doi.org/10.18653/v1/2025.nllp-1.23

[20] He, C., Hu, H., Li, Y., Zhang, H., & Zhang, Q. (2026). A survey of large language models for legal tasks: Progress, prospects, and challenges. *Computer Science Review*, *60*, 100906. https://doi.org/10.1016/j.cosrev.2026.100906

[21] Cho, E., Hoyle, A. M., & Hermstrüwer, Y. (2025). Modeling motivated reasoning in law: Evaluating strategic role conditioning in LLM summarization. In *Proceedings of the Natural Legal Language Processing Workshop 2025*, 68–112. https://doi.org/10.18653/v1/2025.nllp-1.7

[22] Guo, X., Huang, Y., Wei, B., Kuang, K., Wu, Y., Gan, L., . . . , & Dong, X. (2025). Specialized or general AI? A comparative evaluation of LLMs' performance in legal tasks. *Artificial Intelligence and Law*, 1–37. https://doi.org/10.1007/s10506-025-09460-y

[23] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*(9), 1–35. https://doi.org/10.1145/3560815

[24] Yang, W., Ma, S., Lin, Y., & Wei, F. (2025). Towards thinking-optimal scaling of test-time compute for LLM reasoning. In *Advances in Neural Information Processing Systems, 38*, 43605–43631.

[25] Chen, H., Chen, J., & Ding, J. (2021). Data evaluation and enhancement for quality improvement of machine learning. *IEEE Transactions on Reliability*, *70*(2), 831–847. https://doi.org/10.1109/TR.2021.3070863

[26] Zha, D., Bhat, Z. P., Lai, K.-H., Yang, F., Jiang, Z., Zhong, S., & Hu, X. (2025). Data-centric artificial intelligence: A survey. *ACM Computing Surveys*, *57*(5), 1–42. https://doi.org/10.1145/3711118

[27] Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

[28] El Hamdani, R., Bonald, T., Malliaros, F. D., Holzenberger, N., & Suchanek, F. (2024). The factuality of large language models in the legal domain. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 3741–3746. https://doi.org/10.1145/3627673.3679961

[29] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

[30] Colombo, P., Pires, T., Boudiaf, M., Melo, R., Culver, D., Malaboeuf, E., . . . , & Desa, M. (2024). SAULLM-54B & SAULLM-141B: Scaling up domain adaptation for the legal domain. In *Advances in Neural Information Processing Systems*, *37*, 129672–129695. https://doi.org/10.52202/079017-4120

[31] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. *arXiv Preprint:. 2009.03300.*

[32] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

[33] Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11048–11064.

[34] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems, 35*, 22199–22213.

[35] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., . . . , & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 24824–24837. https://dl.acm.org/doi/10.5555/3600270.3602070

[36] Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., . . . , & Chi, E. (2023). Least-to-most prompting enables complex reasoning in large language models. *arXiv. Preprint:2205.10625.*

[37] Bommarito, M. J., II, Katz., M, D., & Detterman, E. M. (2021). LexNLP: Natural language processing and information extraction for legal and regulatory texts. In R. Vogal (Ed.), *Research handbook on big data law* (pp. 216–227). UK: Edward Elgar Publishing. https://doi.org/10.4337/9781788972826.00017

[38] Liga, D., & Robaldo, L. (2023). Fine-tuning GPT-3 for legal rule classification. *Computer Law & Security Review*, *51*, 105864. https://doi.org/10.1016/j.clsr.2023.105864

[39] Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

How to Cite: Wang, S., Pobbathi, L., & Chen, H. (2026). LAMUS: A Large-Scale Corpus for Legal Argument Mining from U.S. Caselaw Using LLMs. *Journal of Computational Law and Legal Technology*. https://doi.org/10.47852/bonviewJCLLT62029649

A Appendix

This appendix provides supplementary results from our experimental evaluation, including the complete ablation grid, stability tests, full model rankings, extended few-shot analysis, per-class performance comparisons, domain shift analysis, and Cohen's kappa interpretation scales. These materials support reproducibility and provide additional detail beyond the main text.

A.1 Complete Ablation Grid

To systematically identify optimal hyperparameters for fine-tuning LLaMA-3-8B on legal argument classification, we conducted a comprehensive grid search across 36 configurations. Table 12 presents accuracy results for all combinations of four learning rates (1e-5, 5e-5, 1e-4, 2e-4), three LoRA ranks (8, 16, 32), and three training durations (1, 3, 5 epochs).

Several patterns emerge from the ablation results. First, learning rate exhibits the strongest influence on performance: at 1 epoch, accuracy ranges from 55.95% (LR = 1e-5, R = 8) to 83.62% (LR = 2e-4, R = 32), a spread of nearly 28 percentage points. Second, the optimal learning rate shifts with training duration—higher learning rates (2e-4) perform best for shorter training (1–3 epochs), while moderate rates (1e-4) excel with longer training (5 epochs), likely because aggressive learning rates cause overfitting when combined with extended training. Third, LoRA rank has minimal impact once learning rate and epochs are optimized, with accuracy varying by less than 1% across ranks at optimal settings. This suggests that rank 8 is sufficient for this classification task, reducing computational overhead without sacrificing accuracy. The best configuration (LR = 1e-4, Rank = 8, Epochs = 5) achieves 85.32% accuracy, matching the result obtained with Rank = 16 at the same settings.

Table 12
Complete ablation grid results for fine-tuned LLaMA-3-8B. All values are test accuracy (%). Best result: 85.32% (LR = 1e-4, Rank = 8, Epochs = 5)

	Epochs = 1			Epochs = 3			Epochs = 5		
LR	R = 8	R = 16	R = 32	R = 8	R = 16	R = 32	R = 8	R = 16	R = 32
1e-5	55.95	66.62	69.24	70.02	75.12	79.60	76.82	79.75	81.76
5e-5	74.96	78.83	81.30	82.23	83.15	83.93	84.23	84.70	85.01
1e-4	79.44	80.68	81.76	83.77	84.54	84.23	85.32	85.32	85.01
2e-4	81.30	82.69	83.62	84.70	85.16	84.39	84.85	83.93	83.31

A.2 Stability Test Results

To assess the reproducibility of our prompting-based results, we conducted stability testing by running chain-of-thought prompting with LLaMA-3-8B across 10 independent trials using different random seeds. Table 13 reports the accuracy achieved in each run.

Table 13
Stability test results for chain-of-thought prompting with LLaMA-3-8B across 10 independent runs. Mean accuracy: 74.71%, standard deviation: 0.56%

Run	Seed	Accuracy (%)
1	142	75.12
2	242	75.43
3	342	74.50
4	442	74.34
5	542	74.96
6	642	74.34
7	742	74.65
8	842	75.58
9	942	73.72
10	1042	74.50
Mean	–	74.71
Std Dev	–	0.56

The results demonstrate high reproducibility, with a standard deviation of only 0.56% across runs. Individual accuracies ranged from 73.72% to 75.58%, representing a spread of less than 2 percentage points. A one-sample t-test confirmed that the mean accuracy (74.71%) significantly exceeds the zero-shot baseline (67.23%) with $p < 0.001$, validating that the performance gains from chain-of-thought prompting are statistically robust rather than artifacts of random variation. This stability supports the reliability of our comparative findings and suggests that practitioners can expect consistent performance when deploying CoT prompting for legal argument classification.

A.3 Complete Model Comparison

Table 14 provides a comprehensive ranking of all model-method combinations evaluated in this study, ordered by accuracy. This complete view contextualizes the relative performance of different approaches and highlights the magnitude of improvement achieved through fine-tuning.

Table 14
Complete results across all models and methods, ranked by accuracy. Baseline (majority class): 61.98%. The Δ column shows improvement over baseline

Model	Method	Acc.	Δ
LLaMA-3-8B	Fine-tuned (best)	85.32%	+23.34
LLaMA-3-8B	Fine-tuned (2e-4)	85.16%	+23.18
LegalBERT	Fine-tuned	81.30%	+19.32
LLaMA-3-8B	Fine-tuned (orig)	80.37%	+18.39
LLaMA-3-8B	Chain-of-thought	75.89%	+13.91
SaulLM-54B	Chain-of-thought	72.80%	+10.82
SaulLM-54B	Zero-shot	67.39%	+5.41
LLaMA-3-8B	Zero-shot	65.38%	+3.40
SaulLM-54B	Few-shot	64.76%	+2.78
law-LLM	Zero-shot	60.12%	–1.86
Qwen3-Thinking	Zero-shot	56.11%	–5.87
Qwen3-Thinking	CoT	54.10%	–7.88
SaulLM-7B	Zero-shot	52.09%	–9.89
Qwen3-Thinking	Few-shot	49.30%	–12.68
LLaMA-3-8B	Few-shot	45.75%	–16.23
SaulLM-7B	CoT	38.02%	–23.96
law-LLM	Few-shot	31.68%	–30.30
law-LLM	CoT	28.75%	–33.23
SaulLM-7B	Few-shot	21.64%	–40.34
Gemini-2.5	All	5.41%	–56.57

The ranking reveals several important patterns. Fine-tuned models occupy the top four positions, with accuracies ranging from 80.37% to 85.32%, demonstrating the substantial advantage of task-specific adaptation. Among prompting approaches, chain-of-thought with LLaMA-3-8B (75.89%) significantly outperforms all other prompted configurations. Notably, few-shot prompting consistently underperforms zero-shot baselines for most models, with LLaMA-3-8B few-shot (45.75%) falling nearly 20 percentage points below its zero-shot result (65.38%). Several smaller models (SaulLM-7B, law-LLM with CoT/few-shot) perform below the majority-class baseline, indicating that neither domain specialization nor advanced prompting compensates for insufficient model capacity. Gemini-2.5 Flash produced near-random outputs (5.41%) due to parsing failures, highlighting the importance of output format compatibility in LLM evaluation.

A.4 Extended Few-Shot Analysis

To investigate the relationship between example count and classification accuracy, we evaluated LLaMA-3-8B using 0–100 few-shot examples. Table 15 reports results for both the full 6-class task and a reduced 5-class variant excluding "Others."

Performance consistently degraded as the number of examples increased. Accuracy declined from 67.23% with zero examples to 53.94% with 100 examples, with similar trends in the 5-class setting. We attribute this degradation to domain mismatch: the curated examples reflected generic legal scenarios, whereas the test data consisted of jurisdiction-specific Texas criminal court language. As

Table 15
Extended few-shot sweep (0–100 examples) for LLaMA-3-8B. Accuracy is reported for both 6-class and 5-class (excluding "Others") settings. Performance degrades monotonically as example count increases

# Ex.	Acc. (6-class)	Acc. (5-class)	Δ ZS
0	67.23%	71.08%	Baseline
5	65.07%	68.79%	–2.16%
10	66.15%	69.93%	–1.08%
20	64.91%	68.63%	–2.32%
40	65.53%	69.28%	–1.70%
60	60.43%	63.89%	–6.80%
80	59.04%	62.42%	–8.19%
100	53.94%	57.03%	–13.29%

additional mismatched examples were introduced, the model increasingly overfit to irrelevant prompt patterns rather than learning generalizable classification criteria. These findings suggest that generic few-shot prompting may harm performance on domain-specific legal classification tasks, making zero-shot or fine-tuned approaches preferable.

A.5 Per-Class Performance

Table 16 compares F1-scores across argument categories for the three best-performing approaches: fine-tuned LLaMA-3-8B, fine-tuned LegalBERT, and chain-of-thought prompting.

Table 16
Per-class F1-scores for top-performing methods. Fine-tuned models achieve more balanced performance across categories, while CoT prompting shows greater variance

Category	LLaMA-FT	LegalBERT	CoT
Facts	0.91	0.92	0.82
Issue	0.73	0.78	0.65
Rule/Law/Holding	0.71	0.46	0.58
Analysis	0.54	0.61	0.48
Conclusion	0.69	0.71	0.62
Others	0.53	0.51	0.41
Weighted Avg	0.80	0.81	0.77
Macro Avg	0.69	0.66	0.59

All methods perform best on "Facts," the majority class, with F1-scores exceeding 0.82. Performance drops substantially for minority classes: "Analysis" and "Others" prove most challenging, with F1-scores below 0.61 across all methods. Interestingly, fine-tuned LLaMA-3-8B achieves the highest macro-average F1 (0.69), indicating more balanced performance across classes despite LegalBERT's marginally higher weighted average (0.81 vs 0.80). LegalBERT particularly struggles with "Rule/Law/Holding" (F1 = 0.46), possibly due to confusion with "Analysis" sentences that also reference legal rules. Chain-of-thought prompting shows the largest performance gap between majority and minority classes, suggesting that explicit reasoning helps most for common patterns but provides less benefit for rare argument types. These results highlight that class imbalance remains a significant challenge for legal argument classification, and future work should explore techniques such as class-weighted loss functions or data augmentation for minority categories.

A.6 Domain Shift Analysis

Table 17 quantifies distributional differences between the Texas Criminal Cases training corpus and the SCOTUS target corpus.

The Texas corpus was heavily dominated by "Facts" (61.9%), reflecting the evidentiary focus of state criminal appeals. In contrast, SCOTUS opinions showed a more balanced distribution, with "Rule/Law/Holding" and "Analysis" each accounting for 27.6% of sentences. The 35.6 percentage-point decrease in "Facts," alongside increases in rule-based and analytical content, highlights substantial differences in argumentative structure between state criminal appeals and Supreme Court opinions. This domain shift helps explain why strong SCOTUS performance requires either fine-tuning on representative target-domain data or prompting strategies that generalize across distributions.

Table 17
Domain shift between Texas Criminal Cases (training) and SCOTUS (target) corpora. Positive shifts indicate categories more prevalent in SCOTUS; negative shifts indicate categories more prevalent in Texas data

Category	Texas (%)	SCOTUS (%)	Shift
Facts	61.9	26.3	–35.6
Issue	4.2	2.1	–2.1
Rule/Law/Holding	5.3	27.6	+22.2
Analysis	14.3	27.6	+13.3
Conclusion	8.9	4.2	–4.7
Others	5.4	12.2	+6.9

A.7 Cohen's Kappa Interpretation

Table 18 presents the standard interpretation scale for Cohen's kappa.

Cohen's kappa ((\kappa)) measures inter-rater agreement while accounting for agreement expected by chance. Our achieved ($\kappa = 0.85$) falls within the "Almost Perfect" range (0.81–1.00), indicating highly consistent annotations between human annotators and the LLM-based labeling pipeline. This exceeds commonly accepted thresholds for corpus construction in NLP and legal annotation research, supporting both the clarity of the annotation guidelines and the reliability of the automated labeling process.

Table 18
Cohen's kappa interpretation scale following Landis and Koch (1977). Our human verification achieved $\kappa = 0.85$, indicating almost perfect agreement

Kappa range	Interpretation
0.81–1.00	Almost Perfect ← Ours: 0.85
0.61–0.80	Substantial
0.41–0.60	Moderate
0.21–0.40	Fair
0.00–0.20	Slight