

September 2025

Do LLMs Truly “Understand” When a Precedent Is Overruled?

Li ZHANG^{a,1}, Jaromir SAVELKA^b, and Kevin ASHLEY^a

^a University of Pittsburgh

^b Carnegie Mellon University

ORCID ID: Li Zhang <https://orcid.org/0000-0003-0375-1793>, Jaromir Savelka <https://orcid.org/0000-0002-3674-5456>, Kevin Ashley <https://orcid.org/0000-0002-3674-5456>

Abstract. Large language models (LLMs) with extended context windows show promise for complex legal reasoning tasks, yet their ability to understand long legal documents remains insufficiently evaluated. Developing long-context benchmarks that capture realistic, high-stakes tasks remains a significant challenge in the field, as most existing evaluations rely on simplified synthetic tasks that fail to represent the complexity of real-world document understanding. Overruling relationships are foundational to common-law doctrine and commonly found in judicial opinions. They provide a focused and important testbed for long-document legal understanding that closely resembles what legal professionals actually do. We present an assessment of state-of-the-art LLMs on identifying overruling relationships from U.S. Supreme Court cases using a dataset of 236 case pairs. Our evaluation reveals three critical limitations: (1) *era sensitivity* – the models show degraded performance on historical cases compared to modern ones, revealing fundamental temporal bias in their training; (2) *shallow reasoning* – models rely on shallow logical heuristics rather than deep legal comprehension; and (3) *context-dependent reasoning failures* – models produce temporally impossible relationships in complex open-ended tasks despite maintaining basic temporal awareness in simple contexts. Our work contributes a benchmark that addresses the critical gap in realistic long-context evaluation, providing an environment that mirrors the complexity and stakes of actual legal reasoning tasks.

Keywords. large language models, legal reasoning, overruled precedents, information retrieval, trustworthy AI

1. Introduction

The recent expansion of the context windows of large language models (LLMs) has opened up possibilities for their application in domains that require understanding of long, complex documents [16]. Transformer-based models [30] have evolved from the original 512-token limit to models capable of processing millions of tokens [15], with key developments including sparse attention mechanisms [11,28] and memory-efficient

¹Corresponding Author: Li Zhang, liz239@pitt.edu

The full dataset can be accessed at <https://github.com/lizhang-AIandLaw/Do-LLMs-Truly-Understand-When-a-Precedent-Is-Overruled>

September 2025

attention algorithms [15]. Alternative approaches have been proposed to extend long-context modeling without changing the model architecture, such as retrieval-augmented generation (RAG) [26,7] and the use of external memory [31].

However, developing effective benchmarks for evaluating long-context understanding remains a significant challenge in the field. Evaluations relying on synthetic tasks, simplified question-answering formats, or artificially constructed scenarios fail to capture the complexity and stakes of real-world document understanding tasks. The legal field, with its reliance on lengthy case files and intricate statutes [23], presents an ideal domain. This paper presents an investigation into the capabilities of long-context LLMs in a particularly consequential legal task: the identification of overruling relationships. In U.S. law, overruling occurs when a later court declares that a prior decision should no longer be followed as binding precedent; we adopt this as the operational definition for this study.

We introduce a new dataset of 236 U.S. Supreme Court case pairs from 1789 to 2025, each consisting of an overruling case and the case(s) it overruled. Using this dataset, we evaluate a selection of long-context LLMs on three distinct tasks designed to probe the limits of their comprehension. We provide the models with the full texts of the overruling case which explicitly states that it overruled the specific case(s), forcing them to navigate a long context to identify the overruling relationship between them.

This task is challenging even when judicial opinions explicitly use terms like “overruled.” The underlying reasoning process is distributed across pages of text, requiring readers to integrate complex legal arguments across sections into a coherent logical chain. Judicial opinions are filled with specialized legal terminology and subtle nuances that require deep contextual understanding [1]. Our work addresses the critical gap in long-context evaluation by providing a benchmark that captures the realistic complexity of legal reasoning while maintaining the scientific rigor needed for meaningful model assessment.

2. Related Work

2.1. Long-Context Language Model Evaluation

The evaluation of long-context models has led to the development of specialized benchmarks. LongBench [2] introduced a comprehensive evaluation suite covering multiple languages and domains, while LongEval [13] and GSM-Infinite [36] focused specifically on the degradation of performance as context length increases. The SCROLLS benchmark [25] provided domain-specific evaluation across various long-document tasks. LongDocURL [5] and MMLONGBENCH-DOC [17] built multimodal long document benchmarks integrating understanding and reasoning. These benchmarks have revealed that performance often degrades significantly when information is embedded deep within long contexts, particularly for tasks requiring complex reasoning rather than simple information retrieval.

Our work extends this line of research by introducing a novel legal reasoning task that specifically tests models’ ability to identify overruling relationships across lengthy judicial opinions. Unlike previous benchmarks that focus on general document understanding, our task requires temporal reasoning, hierarchical relationship identification, and the integration of complex legal arguments distributed across extended text.

September 2025

2.2. LLMs Application and Evaluation in Legal Domain

The application of LLMs in the legal domain has seen significant growth, driven by the promise of automating complex legal tasks [24,9] and improving access to legal information [27]. Several legal evaluation benchmarks have emerged to assess LLM performance in legal tasks. The LexGLUE benchmark [3] provides a unified evaluation framework covering multiple legal NLP tasks including case outcome prediction, legal text classification, and legal question answering. The CUAD dataset [12] focuses specifically on contract understanding and analysis, testing models' ability to extract and reason about complex contractual terms. Other benchmarks [10,34,18,21] offer evaluation methods covering reasoning, information extraction, and document analysis tasks across various legal domains. These benchmarks have demonstrated that while LLMs show promising results on some legal tasks, they struggle with complex reasoning that requires deep understanding of legal principles.

Our work contributes to this body of research by introducing novel tasks that specifically test LLMs' ability to identify overruling relationships. Previous evaluations also mainly focus on multiple choice QA and classification tasks while our study tests the models' ability to navigate real-world full-context judicial opinions by designing three tasks with a mixture of open-ended and closed-ended questions.

3. Dataset

To evaluate the performance of LLMs on the tasks of identifying overruled precedents, we created a new dataset of 236 U.S. Supreme Court case pairs in the period from 1789 to 2025. Each pair consists of a case that was explicitly overruled and the subsequent decision(s) that overruled it.

The foundation of our dataset is a resource maintained by the Constitution Annotated [4] consisting of 236 U.S. Supreme Court case pairs. The compilation process involved a comprehensive search of the Lexis database for all Supreme Court decisions containing the word “overrule” in the headnotes, syllabus, or opinion. A decision is included only when a majority of the Supreme Court has explicitly stated that it is being overruled. This strict criterion avoids the ambiguity inherent in legal commentary and ensures that each case in our dataset has been definitively identified as overruled by the Court itself. Conversely, cases that are merely distinguished, limited, or discredited without an explicit overruling are excluded. This conservative approach ensures the dataset's reliability. The results were then manually reviewed to confirm the Court's intent and to ensure that each case met the strict inclusion criteria.

For each overruling case of the 236 case pairs, we retrieved the full opinion texts from CourtListener [29], including majority and lead opinions as well as combined, concurring, and dissenting opinions. We then cleaned the raw downloads to retain only the textual content and conducted a manual review to verify content quality and completeness. The resulting curated corpus constitutes the dataset used in this study.

Figure 1 illustrates the distribution of word counts across our dataset. The 236 overruling cases range from the shortest case with 705 words to the longest with 116,455 words. The median case length of 17,607 words and mean case length of 19,368 suggest that most cases in our dataset require efforts to process and understand, making them ideal for testing the limits of LLM long-context comprehension.

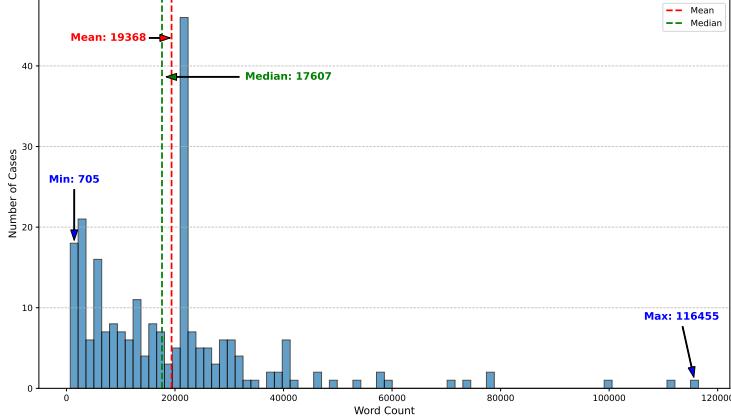


Figure 1. Word Count Distribution of 236 Overruling Cases in Our Dataset.

4. Experiment Design

To probe the capabilities of LLMs’ long-context understanding in the legal domain, we designed three tasks centered on identifying overruling relationships.

4.1. Task 1: Open-Ended Identification

The first task poses an open-ended question to LLMs: “Which case(s) was overruled by [the overruling case, e.g., *“Hohn v. United States*, 524 U.S. 236 (1998)”]? Here is the related context [case text of the overruling case].” In this prompt, we use the overruling case as the context and ask the LLM to identify the case(s) that were overruled by the overruling case.

Rationale This task simulates a realistic legal research query. To evaluate the open-ended responses, we employed GPT-4o as an automated judge, using a few-shot prompt to ensure consistent and accurate assessment against the ground truth. The core prompt was designed as “Do [the model’s response, e.g., *“House v. Mayo”*] and the ground truth [the overruled case(s), e.g., *“House v. Mayo*, 324 U.S. 42 (1945)] refer to the same case?”

The development of the few-shot examples involved two rounds of iterative testing. First, we compared GPT-4o’s zero-shot judgment results against our ground truth annotations and analyzed the discrepancies to identify patterns in misclassification. Based on these findings, we refined the prompt by incorporating carefully selected examples that addressed the common failure modes. In the second round, we conducted a validation experiment where 10% of the responses were randomly selected for human expert review, demonstrating that GPT-4o’s accuracy as a judge under the few-shot prompting approach was 98% consistent with human expert assessments.

4.2. Task 2: Closed-Ended Verification

The second task is a verification query: “Did [the overruling case, e.g., *“Hohn v. United States*, 524 U.S. 236 (1998)”] overrule [the overruled case, e.g., *“House v. Mayo*, 324

September 2025

U.S. 42 (1945)]? Here is the related context [case text of the overruling case].” With this prompt, the model must answer “true,” “false,” or “unknown.” The correct answer is always “true.”

Rationale This task reduces the problem from open-ended QA to closed-ended verification. It allows us to isolate the model’s ability to confirm or deny a specific legal relationship when presented with all necessary context, testing its comprehension and reasoning over long distances within the text.

4.3. Task 3: Reversed Closed-Ended Verification

The third task presents a logically flawed question: “Did [the overruled case, e.g., “*House v. Mayo*, 324 U.S. 42 (1945)] overrule [the overruling case, e.g., “*Hohn v. United States*, 524 U.S. 236 (1998)]”? Here is the related context [case text of the overruling case].” With this prompt, the model must answer “true,” “false,” or “unknown.” The year of each case is provided in the context. The correct answer is always “false,” as a precedent can only be overruled by a later decision.

Rationale This task serves as a control to test whether failures in the other tasks stem from a fundamental misunderstanding of temporal logic or from difficulties in understanding the long context. A model that understands chronology should consistently answer “false”.

4.4. Models

We evaluated five state-of-the-art LLMs known for their strong performance on long-context benchmarks. At the time of our experiments, the Gemini-Pro [8], Gemini-Flash, Qwen3 [32] represented the top 3 of the LongBench v2 leaderboard [2], and were selected to ensure our evaluation reflected the current state of the art. We also included GPT-5 [20] and Gemini-Flash-Lite to broaden our assessment. The models and their context window length (in tokens) are:

- Qwen3-235B-A22B-Thinking-2507 (Qwen3) - Context Length: 262,144
- Gemini-2.5-flash-lite-preview-06-17 (Gemini-Flash-Lite) - Context Length: 1,000,000
- Gemini-2.5-flash-preview-05-20 (Gemini-Flash) - Context Length: 1,000,000
- Gemini-2.5-pro-preview-06-05 (Gemini-Pro) - Context Length: 1,000,000
- GPT-5-2025-08-07 (GPT-5) - Context Length: 128,000

4.5. Experimental Setup

All experiments were conducted with a temperature of 0.1, a top-p of 0.5, and a structured json output format, to encourage deterministic responses. For GPT-5 model, we use the default parameters as set by OpenAI, as the relevant parameter controls are not publicly exposed.

5. Empirical Results

The performance of the selected LLMs is summarized in Table 1.

September 2025

Table 1. Model Performance Across Tasks (Accuracy in %, higher is better; \pm denotes 95% confidence intervals, $n = 236$).

Model	Task 1	Task 2	Task 3
Qwen3	2.12% \pm 1.84%	1.27% \pm 1.42%	97.46% \pm 2.02%
Gemini-Flash-Lite	0.85% \pm 1.01%	11.86% \pm 4.08%	91.95% \pm 3.52%
Gemini-Flash	33.47% \pm 6.03%	6.36% \pm 3.12%	62.71% \pm 6.12%
Gemini-Pro	73.31% \pm 5.65%	42.37% \pm 6.25%	100% \pm 0%
GPT-5	71.19% \pm 5.78%	75.42% \pm 5.47%	100% \pm 0%

5.1. Task 1: Open-Ended Identification

In the open-ended identification task, the models struggled to correctly identify the overruled case. The best-performing model, Gemini-Pro, achieved an accuracy of 73.31% (173 out of 236), closely followed by GPT-5 with 71.19% (168 out of 236). The other models performed worse. This suggests that even state-of-the-art models have difficulty with understanding the overruling relationship in a long legal context.

There are three patterns in the errors. First, when an overruling case overturns multiple precedents, models often extract only one of them, yielding a partially correct answer. Second, models sometimes confuse cases that are merely mentioned in the text with the case actually overruled—we refer to this as *confusion* errors, where models incorrectly identify a case that appears in the text but is not the overruled precedent. Third, models hallucinate a case that is not mentioned in the opinion at all—we define *hallucination* as the generation of case names or citations that do not exist in the provided context; for example, in Gemini-Flash-Lite’s response to *Garland v. Washington*, the model produced *Miles v. Graham*, which is neither cited in the text nor overruled by *Garland*. In the calculation of the accuracy, we only count the cases where the model correctly identified all overruled cases. Table 2 provides illustrative examples of these patterns.

Table 2. Examples of Model Responses and Error Patterns for Task 1

Overruling Case	Overruled Case	Model Response	Type
<i>Hohn v. United States</i> , 524 U.S. 236 (1998)	<i>House v. Mayo</i> , 324 U.S. 42 (1945)	<i>House v. Mayo</i> , 324 U.S. 42 (1945)	Correct
<i>Payne v. Tennessee</i> , 501 U.S. 808 (1991)	<i>South Carolina v. Gathers</i> , 490 U.S. 805 (1989) & <i>Booth v. Maryland</i> , 482 U.S. 496 (1987)	<i>South Carolina v. Gathers</i> , 490 U.S. 805 (1989)	Missing Case
<i>Garland v. Washington</i> , 232 U.S. 642 (1914)	<i>Crain v. United States</i> , 162 U.S. 625 (1896)	<i>Rogers v. Peck</i> , 199 U.S. 425, 435 (1905)	Confusion
<i>Garland v. Washington</i> , 232 U.S. 642 (1914)	<i>Crain v. United States</i> , 162 U.S. 625 (1896)	<i>Miles v. Graham</i> , 268 U.S. 501 (1925)	Hallucination

5.2. Task 2 & 3: Closed-Ended Verification

The closed-ended verification of Task 2 yielded more varied results. GPT-5 emerged as the top performer in this task, achieving an accuracy of 75.42% (178 out of 236), significantly outperforming Gemini-Pro which achieved 42.37% (100 out of 236). GPT-5 also showed a much lower rate of abstention, answering “unknown” in only 3 cases

compared to Gemini-Pro’s 131 abstentions. However, this low abstention rate comes with a significant cost: GPT-5 incorrectly labeled 55 cases as “False,” representing a substantial number of false negatives.

While GPT-5’s low abstention rate might suggest higher confidence in its reasoning capabilities, it could also be interpreted as a tendency toward overconfident incorrectness. This trade-off highlights the need for more sophisticated uncertainty quantification mechanisms that can distinguish between genuine epistemic uncertainty and mere comprehension gaps. The other models showed a significant drop in performance, with Gemini-Flash answering “unknown” in the majority of cases. This suggests that even when the task is simplified to verification, most models struggle to confidently identify the correct legal relationship.

The reversed closed-ended verification of Task 3 showed a marked improvement in performance across all models. The accuracy was higher than in Task 2. GPT-5 and Gemini-Pro achieved high accuracy (100%), correctly identifying the logical impossibility in all cases. Qwen3 also performed well with 97.5% accuracy.

5.3. Discussion

5.3.1. Era Sensitivity

To probe temporal robustness in the open-ended identification task (Task 1), we stratified cases by overruling case decision year (the earliest overruling case happened in 1810) and measured models’ accuracy across the fixed historical intervals that reflect changes in case density and legal language evolution. There are 17 case pairs from 1810–1881, 65 case pairs from 1882–1953, and 154 case pairs from 1954–2025. This quantitative view in Table 3 aligns with our qualitative observations and clarifies where failures concentrate.

Table 3. Model Performance by Historical Intervals in Task 1 (Accuracy in %, higher is better).

Model	1810–1881	1882–1953	1954–2025
Qwen3	0.00%	1.54%	2.60%
Gemini-Flash-Lite	0.00%	0.00%	1.30%
Gemini-Flash	17.65%	24.62%	38.96%
Gemini-Pro	64.71%	72.31%	74.68%
GPT-5	35.29%	53.85%	82.47%

The temporal stratification reveals era sensitivity across all models, with performance degrading as we move backward in time. In the earliest period (1810–1881), the performance gap is stark: Qwen3 and Gemini-Flash-Lite achieve 0% accuracy, completely failing to identify any overruling relationships in this historical interval. The best-performing model in this period, Gemini-Pro, manages only 64.71% accuracy, while GPT-5 struggles at 35.29%. This performance drop in the earliest cases suggests that models trained primarily on modern legal texts lack the linguistic and conceptual frameworks needed to parse 19th-century judicial opinions, which often employ archaic legal terminology, different citation conventions, and distinct argumentative structures [19].

The middle period (1882–1953) shows a moderate improvement. Here, Gemini-Pro maintains its strong performance at 72.31% accuracy, demonstrating remarkable temporal consistency. GPT-5 shows significant improvement to 53.85%, while Gemini-Flash

reaches 24.62%. However, the weaker models continue to struggle: Qwen3 achieves only 1.54% accuracy, and Gemini-Flash-Lite reaches 0.00%. This intermediate period represents a transitional phase where legal language begins to modernize but still retains many historical characteristics that challenge contemporary models. The performance improvements suggest that models can adapt to some degree of historical variation, but their capabilities remain fundamentally limited when dealing with pre-modern legal discourse.

The modern period (1954–2025) represents the models’ comfort zone, with significantly improved performance across the board. GPT-5 achieves its highest accuracy at 82.47%, significantly outperforming Gemini-Pro’s 74.68%. Gemini-Flash shows improvement to 38.96% while remaining limited. This improvement in the modern era reflects several factors: the models’ training data heavily emphasizes contemporary legal texts and the legal reasoning patterns align more closely with what the models have learned during training [6]. The fact that even the best-performing models achieve only around 80% accuracy underscores the fundamental challenge of legal reasoning tasks, while the precipitous drop in performance for earlier periods quantifies the temporal bias that underlies the open-ended reasoning failures we observed in our experiments.

5.3.2. Hesitation in Verification vs. Confidence in Refutation

The high number of “unknown” responses in Task 2 is revealing, as shown in Figure 2. It suggests that most models are not simply guessing, but are actively assessing their own uncertainty. When the models are unable to reason and understand the case relationship embedded in the text, they are hesitant to make a definitive judgment. This is a desirable trait in a legal AI system, as it is preferable for a model to admit its own ignorance rather than to provide a confident but incorrect answer. However, the high rate of abstention in this task suggests that most models’ threshold for certainty is set too high, preventing them from making correct judgments even when the evidence is strong. GPT-5’s low abstention rate (1.27%, 3 out of 236) stands in stark contrast to this pattern, suggesting a different approach to uncertainty assessment.

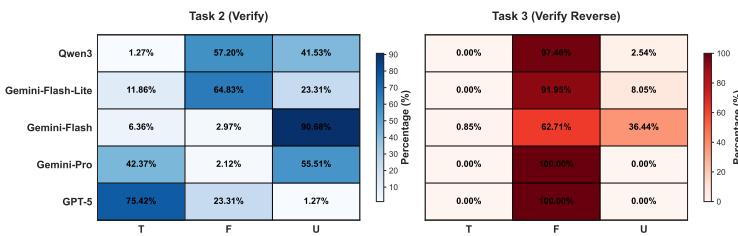


Figure 2. Confusion Matrix for Verification Tasks. T, F, and U stand for True, False, and Unknown respectively.

A contrast emerges when comparing the results of Task 2 and Task 3. In Task 2, when asked to verify a correct overruling relationship, the models frequently defaulted to “unknown,” even when the answer was clearly stated in the text. In contrast, when presented with the temporally impossible proposition in Task 3, the same models answered with high confidence and accuracy, with GPT-5 and Gemini-Pro achieving 100% accuracy and Qwen3 achieving 97.46%.

September 2025

This pattern of behavior reveals a fundamental limitation in the models' approach to legal reasoning. The contrast between Task 2 and Task 3 performance demonstrates that models rely on shallow logical heuristics rather than deep legal comprehension. In Task 3, models can successfully apply simple temporal logic—recognizing that an earlier case cannot overrule a later one—without actually understanding the complex legal reasoning embedded in the judicial opinions. This high performance is achieved through basic chronological awareness rather than genuine comprehension of legal arguments.

5.3.3. Temporal Reasoning

Our analysis also reveals an interesting pattern in temporal reasoning across different task contexts. In the open-ended identification task (Task 1), we observed what we term “context-dependent temporal reasoning failures,” as illustrated in Table 4. These failures show that models can create temporally impossible relationships, such as suggesting that a 1914 case overruled a 1925 case, even when the case names include the years. Such responses indicate a breakdown in temporal causality, where a case from an earlier year is incorrectly identified as overruling a case from a later year.

Table 4. Examples of Context-Dependent Temporal Reasoning Failures from Task 1. LLMs generated logically impossible scenarios where a subsequent case was overruled by a preceding one.

Case Relationship / Query	Year
<i>Query Case (Overruling Case):</i>	
Garland v. Washington	1914
<i>Model's Response (Overruled Case):</i>	
Miles v. Graham	1925

However, when we examine the results from Task 3 (the reversed verification task), a different picture emerges. As shown in Table 1, models demonstrate remarkably high accuracy in identifying temporally impossible propositions. GPT-5 and Gemini-Pro achieved 100% accuracy, while Qwen3 achieved 97.46% accuracy in correctly rejecting statements like “Did *Booth v. Maryland* (1987) overrule *Payne v. Tennessee* (1991)?”—a clear temporal impossibility.

This contrast suggests that the models are not fundamentally lacking in temporal reasoning capabilities or awareness of legal case chronology. Rather, the issue lies in the complexity of the reasoning required in different contexts. In Task 3, the models are presented with a straightforward logical contradiction that they can easily identify using basic temporal logic. In contrast, Task 1 requires more sophisticated reasoning—the models must understand complex legal relationships, extract relevant information from dense legal texts, and apply temporal logic in an open-ended context.

6. Limitations and Future Work

Our study has several limitations. First, our dataset focuses exclusively on U.S. Supreme Court cases, which may not generalize to other jurisdictions or legal systems with different precedent structures. Second, we only examined cases with explicit overruling language, excluding more nuanced legal developments like distinguishing, limiting, or discrediting precedents. Third, our evaluation relies on automated assessment using GPT-

September 2025

4o, which, while validated against human experts, may introduce its own biases. Finally, we tested a limited set of models at a specific point in time, and rapid advances in LLM technology may render some findings obsolete.

Future work should address these limitations by expanding to international legal systems, incorporating more subtle legal relationships, and developing more sophisticated evaluation metrics. Additionally, research should explore architectural improvements for long-context understanding, such as legal domain-specific pre-training, better uncertainty quantification, and the potential of RAG systems [14,22] to enhance LLMs’ long-context understanding capabilities. Exploring agent-based approaches [35,33] for enhancing LLMs’ long-text capabilities could also lead to more consistent and reliable legal AI systems. The development of legal reasoning benchmarks that test deeper comprehension would further advance the field, along with more detailed quantitative analysis of error patterns across different models to strengthen qualitative insights into specific failure modes.

7. Conclusion

This paper presents an evaluation of state-of-the-art LLMs on the legal reasoning task of identifying overruling relationships from U.S. Supreme Court cases. Our investigation reveals three fundamental limitations that challenge the practical deployment of these models in legal contexts.

First, we demonstrate that models exhibit era sensitivity. The temporal stratification reveals era sensitivity across all models in our evaluation, with performance degrading as we move backward in time. This temporal bias reveals that extended context windows alone cannot overcome the fundamental challenge of understanding legal language across different historical periods, suggesting that models lack the linguistic and conceptual frameworks needed to parse judicial opinions from earlier eras that employ archaic legal terminology, different citation conventions, and distinct argumentative structures.

Second, our analysis reveals the models’ reliance on shallow logical heuristics rather than deep legal comprehension. Models can successfully apply simple temporal logic—such as recognizing that an earlier case cannot overrule a later one—without actually understanding the complex legal reasoning embedded in the judicial opinions. The high performance on Task 3 is achieved through basic chronological awareness rather than genuine comprehension of legal arguments, highlighting a fundamental gap between surface-level logical reasoning and deep legal understanding.

Third, we identify context-dependent temporal reasoning failures that highlight the complexity of legal reasoning tasks. Models can correctly identify simple temporal contradictions but fail to maintain temporal consistency when processing complex, open-ended legal relationships. This finding indicates that the challenge lies not in basic logical reasoning, but in integrating temporal awareness with deep legal understanding under cognitive load.

The empirical results underscore a critical gap between current LLM capabilities and legal practice requirements. Even the best-performing model achieved only 73.31% accuracy on the core identification task, falling short of the reliability standards needed for legal applications. This performance gap is particularly concerning given that our dataset focuses on cases with explicit overruling language, representing a straightforward scenario for legal relationship identification.

References

- [1] Kevin D Ashley. *Artificial intelligence and legal analytics: new tools for law practice in the digital age*. Cambridge University Press, 2017.
- [2] Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, et al. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *arXiv preprint arXiv:2412.15204*, 2024.
- [3] Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. Lexglue: A benchmark dataset for legal language understanding in english. *arXiv preprint arXiv:2110.00976*, 2021.
- [4] Congressional Research Service. Constitution of the united states: Analysis and interpretation. <https://constitution.congress.gov>.
- [5] Chao Deng, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, et al. Longdocurl: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. *arXiv preprint arXiv:2412.18424*, 2024.
- [6] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.
- [7] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1), 2023.
- [8] Google. Model cards. <https://modelcards.withgoogle.com/model-cards>, 2025. Accessed on 2025-08-26.
- [9] Morgan Gray, Li Zhang, and Kevin D Ashley. Generating case-based legal arguments with llms. In *Proceedings of the 2025 Symposium on Computer Science and Law*, pages 160–168, 2025.
- [10] Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in neural information processing systems*, 36:44123–44279, 2023.
- [11] Chi Han, Qifan Wang, Hao Peng, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. Lm-infinite: Zero-shot extreme length generalization for large language models. *arXiv preprint arXiv:2308.16137*, 2023.
- [12] Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268*, 2021.
- [13] Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. Longeval: Guidelines for human evaluation of faithfulness in long-form summarization. *arXiv preprint arXiv:2301.13298*, 2023.
- [14] Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. Retrieval augmented generation or long-context llms? a comprehensive study and hybrid approach. *arXiv preprint arXiv:2407.16833*, 2024.
- [15] Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024.
- [16] Jiaheng Liu, Dawei Zhu, Zhiqi Bai, Yancheng He, Huanxuan Liao, Haoran Que, Zekun Wang, Chenchen Zhang, Ge Zhang, Jiebin Zhang, et al. A comprehensive survey on long context language modeling. *arXiv preprint arXiv:2503.17407*, 2025.
- [17] Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems*, 37:95963–96010, 2024.
- [18] Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. Hallucination-free? assessing the reliability of leading ai legal research tools. *Journal of Empirical Legal Studies*, 22(2):216–242, 2025.
- [19] David Mellinkoff. *The language of the law*. Wipf and Stock Publishers, 2004.
- [20] OpenAI. Gpt-5 system card. <https://cdn.openai.com/gpt-5-system-card.pdf>, 2025. Accessed on 2025-08-26.
- [21] Nicholas Pipitone and Ghita Hour Alami. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. *arXiv preprint arXiv:2408.10343*, 2024.

September 2025

- [22] Zehan Qi, Rongwu Xu, Zhijiang Guo, Cunxiang Wang, Hao Zhang, and Wei Xu. Long²rag: Evaluating long-context & long-form retrieval-augmented generation with key point recall. *arXiv preprint arXiv:2410.23000*, 2024.
- [23] John B Ruhl. Law's complexity: a primer. *Ga. St. UL Rev.*, 24:885, 2007.
- [24] Jaromir Savelka and Kevin D Ashley. The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts. *Frontiers in Artificial Intelligence*, 6:1279794, 2023.
- [25] Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhao Xiong, Mor Geva, Jonathan Berant, et al. Scrolls: Standardized comparison over long language sequences. *arXiv preprint arXiv:2201.03533*, 2022.
- [26] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*, 2023.
- [27] Francesco Sovrano, Kevin Ashley, Peter Leonid Brusilovsky, and Fabio Vitali. How to improve the explanatory power of an intelligent textbook: a case study in legal writing. *International Journal of Artificial Intelligence in Education*, pages 1–35, 2024.
- [28] Jiaming Tang, Yilong Zhao, Kan Zhu, Guangxuan Xiao, Baris Kasikci, and Song Han. Quest: Query-aware sparsity for efficient long-context llm inference. *arXiv preprint arXiv:2406.10774*, 2024.
- [29] The Free Law Project. RECAP archive, 2020.
- [30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [31] Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. *Advances in Neural Information Processing Systems*, 36:74530–74543, 2023.
- [32] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengan Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- [33] Li Zhang and Kevin D Ashley. Mitigating manipulation and enhancing persuasion: A reflective multi-agent approach for legal argument generation. *LCIC @ ICAIL 2025*, 2025.
- [34] Li Zhang, Morgan Gray, Jaromir Savelka, and Kevin D Ashley. Measuring faithfulness and abstention: An automated pipeline for evaluating llm-generated 3-ply case-based legal arguments. *ASAIL @ ICAIL*, 2025.
- [35] Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Arik. Chain of agents: Large language models collaborating on long-context tasks. *Advances in Neural Information Processing Systems*, 37:132208–132237, 2024.
- [36] Yang Zhou, Hongyi Liu, Zhuoming Chen, Yuandong Tian, and Beidi Chen. Gsm-infinite: How do your llms behave over infinitely increasing context length and reasoning complexity? *arXiv preprint arXiv:2502.05252*, 2025.