Beyond the Haystack: Sensitivity to Context in Reference Recall

Karthik Srikumar[3] **Keshav Karthik**[1] **Advaith Renjith**[1] **Eric Xia**[2]

[1]Algoverse AI Research [2]Brown University [3]South Windsor High School

Abstract

Long-context benchmarks such as Needle in a Haystack (NIAH) are often cited as evidence of strong recall abilities in large language models (LLMs). However, these evaluations typically remove target phrases from contextual information, making their validity in naturalistic recall tasks uncertain. Using a dataset of post-cutoff U.S. court opinions, we assess recall across standard legal texts and systematically altered inputs, including a globally shuffled condition inspired by NIAH-style benchmarks. Our results reveal a distinct J-shaped performance curve: accuracy drops with localized shuffling, but rises to an overall maximum in globally shuffled texts. This suggests that models employ context-sensitive retrieval, relying on surrounding information when possible and on blind recall in other conditions. By finding models most accurately perform recall in globally shuffled contexts, we highlight a need for context-aware evaluation within reference-critical domains.

1 Introduction

Language model (LM) benchmarks (Gao et al., 2023) include evaluations which measure the long-context recall performance of models. This has been done primarily through variations of the Needle In A Haystack format, in which a short string of text (the needle) is inserted in a larger context window (the haystack). The extent to which NIAH recall scores transfer to specialized domains such as law remains an open question.

Existing benchmarks often conflate retrieval with downstream reasoning, and lack systematic approaches to avoiding data contamination (LeP, 2024; Chalkidis et al., 2021; Guha et al., 2023). We present a framework for closed-domain legal recall benchmarks using court opinions published after model training cutoffs to test retrieval from unseen texts. Our matched needle task requires models to recover real legal references and facts, revealing a consistent performance gap between standard NIAH tasks and true legal recall, even after controlling for context by using shuffled texts.

We find that models likely do not use recall to succeed in standard long-context benchmarks; instead, they process inputs via sentence-to-sentence cues on texts. We observe a J-shaped curve in model performance as shuffle window increases, with performance decreasing for local shuffles and increasing to a maximum for global shuffles. Furthermore, we perform position ablations by inserting the needle in different locations and, finding no correlation between position and accuracy, show that positional bias does not affect performance. By emphasizing this distinction between reading and recall ability that is often conflated by standard long-context benchmarks, we highlight the need for new long-context benchmarks that properly isolate recall.

2 Related Work

Hallucination studies. Language model hallucinations, where models provide plausible but factually incorrect answers to user queries, is a problem of particular importance in the legal domain, where arguments depend on the verifiability of prior work. Existing literature has investigated causes, types, and strategies for reducing hallucinations in both open and closed domains (Dahl et al., 2024; Hu et al., 2025; Li, 2023).

Legal Benchmarks. Having benchmarks that evaluate legal understanding and citation retrieval is crucial to assess how capable models are of being deployed in the legal field. These include LexGLUE (Chalkidis et al., 2021), LePaRD (LeP, 2024), and LegalBench (Guha et al., 2023; Houir Alami et al., 2024; Zheng et al., 2025).However, these works often blend retrieval with reasoning components and do not ensure evaluation

Proceedings of the Natural Legal Language Processing Workshop 2025, pages 48–53

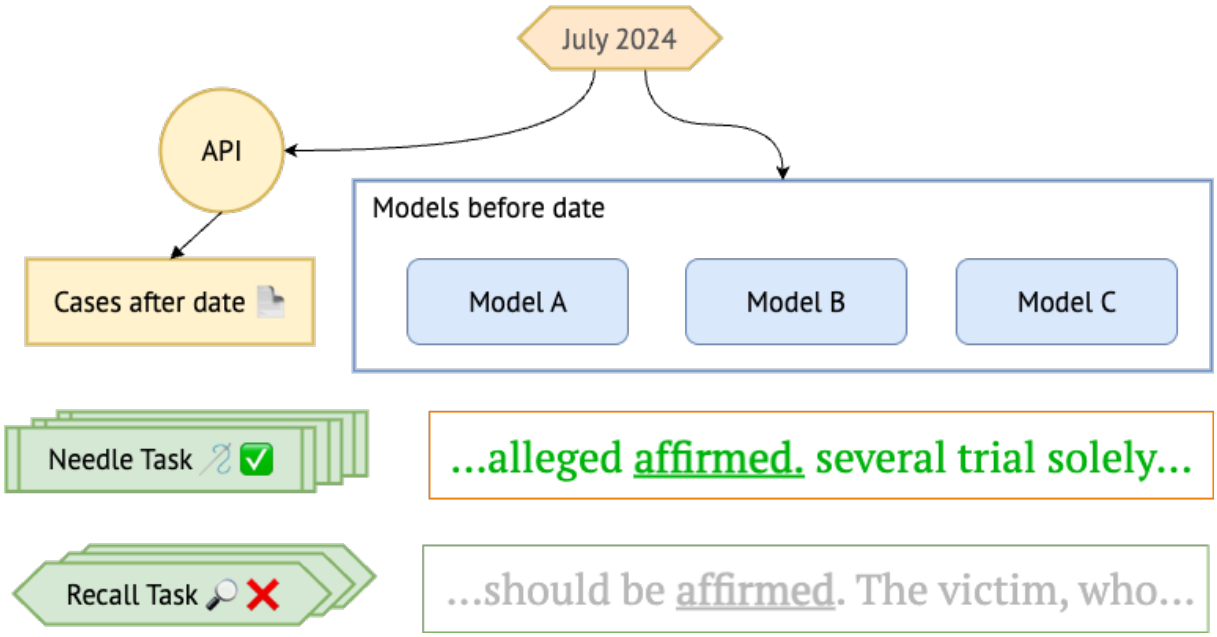

Figure 1: In a closed-domain setting, legal reference recall is consistently lower than distributionally identical needle-in-a-haystack tasks, suggesting context-specific information impacts legal reference recall.

independence from training data.

Long Context Recall. Many recall benchmarks test token extraction, these include Needle-in-a-Haystack (NIAH) (Gao et al., 2023) and its variants, including NeedleChain (Moon and Lim, 2025) and recall benchmarks in reasoning and recall-based applications (Yu et al., 2025; Wang, 2025; Gupta et al., 2024; Qiu et al., 2025). However, these works often permit preexisting mentions of the needle fragment in training data (Chen et al., 2023; Mamakas et al., 2022). There also exist several studies which conduct recall within a specialized domain. Blair-Stanek et al. (2024) specifically tests model retrieval from legal contexts, while Fan et al. (2024) assess long-context recall in the medical domain. However, these works omit causal analysis and do not guarantee a lack of data contamination.

3 Approach

3.1 Evaluation Framework

All case texts were sourced from Courtlistener, an online database of American case law containing 1.75 million legal decisions. To ensure that the models' performance is independent of existing knowledge on the provided opinions, we filtered for cases which were made publicly available after a specific cutoff date, for which we selected July 1st, 2024.

We evaluated five state-of-the-art language models with knowledge cutoffs preceding our document collection period: GPT-4o, Gemini 1.5 Pro, Claude Sonnet 3.5, DeepSeek-V3, and Llama-3.1-8b-instruct (OpenAI, 2024; Google, 2024; Anthropic, 2024; DeepSeek-AI, 2024; Meta, 2024). The models tested span parameter scales from 8B to an estimated 1.8T parameters and incorporate diverse architectural approaches, providing a representative sample of current language models.

3.2 Retrieval Evaluation and Ablations

All tests on a document were performed with shuffled versions of the context window, such that each test was distributionally identical. To understand LLM recall as it relates to context information, we ran several types of shuffle tests, including triad shuffles, sentence-level shuffles, paragraph-level shuffles, and global shuffles. The global shuffle corresponds to the Needle In Haystack benchmark, which typically involves contextually irrelevant needle insertions (Gao et al., 2023).

Additionally, we ran ablation tests using needle position to determine the extent to which positional bias influences accuracy. While running neighborhood shuffles to vary contextual information, we evaluated the position within the text of the needle word, and compared retrieval accuracies at each position.

Contextual Information. The shuffles differ in the amount of contextual information provided. For needle word w, context c, and co-occurrence prob-

Context Type	Needle Ex.	Prompt	Case Count
Decision	"affirm"	In one word, was this case affirmed or reversed? {case text}	62
Author	"Karen Moore"	Is {author name} mentioned or identifiable in this case? Answer only Yes or No. Here is a legal document: {case text}	62
Citation	"2015 UT 45"	Here is a legal decision describing a ruling, does the briefing cite {citation}? Answer only Yes or No.: {case text}.	100

Table 1: Summary of the reference-based tasks used in closed-domain recall. For each task, we test both the standard context and a shuffled NIAH variant. For the citation task, we constructed a balanced dataset of 50 real citations and 50 fakes. Fakes were generated by randomly selecting digits from real citations and permuting two digit positions, creating in-distribution but invalid citation references.

ability $P(w, c)$, the pointwise mutual information for a particular word (Resnik, 1992) is given as

$$\text{PMI}(w, c) = \log_2 \frac{P(w, c)}{P(w)P(c)}$$

For a particular needle w, the consequent information given by any context c is then

$$\text{PMI}_{\text{standard}}(w, c) = n$$

for some baseline value n, while for the shuffled context it is

$$\text{PMI}_{\text{NIAH}}(w, c) = 0$$

Because all words have equal likelihood of appearing in any context, $P(w, c) = P(w)P(c)$ in this case so that the mutual information is zero.

3.3 Recall Tasks

We use a subset of tasks that focus on specific factual information within legal settings. Our first task, Decision, tests whether the models can accurately identify a single instance of a court decision, which is either 'affirmed' or 'reversed'. To add more breadth to our evaluation, we introduced our second task, Author, which tests the model's ability to identify a single-occurrence author name from the briefing. Finally, our third task, Citation, evaluates the models on how reliably they can identify a single-occurrence citation from the text.

4 Results

Our results show that model recall follows a J-shaped trajectory as contextual information is disrupted. As shown in Figure 3, interfering with local semantics via small shuffle windows significantly impairs performance, hindering the model's "reading" process. However, completely destroying the context via a global shuffle forces the model into its "recall" mode, boosting accuracy to its peak. Ultimately, every model tested performed much worse in the standard setting compared to the contextually uninformative global shuffle (NIAH setting), suggesting that such benchmarks can be unreliable.

One potential counterfactual to consider is that position alone can explain the drops in accuracy. We eliminate this possibility through needle position ablation tests. As seen in the appendix, these tests show little correlation between position and accuracy and indicate that position alone cannot explain the gap.

The performance gap between standard and NIAH contexts is exacerbated by document length. While a model's NIAH performance is stable, its ability to recall from a coherent legal text degrades sharply as the text gets longer. As seen in Figure 4, the error rate in a standard context increases 3.1x faster than in a NIAH task. This widening gap demonstrates that high NIAH scores are unreliable predictors for performance, especially on long domain-specific documents.

Specific accuracy scores for each model across the three key conditions—Standard (reading), Local Shuffle (disrupted reading), and NIAH (recall)—are detailed in Table 2.

Model	Std.	Local Shuffle	Global Shuffle
GPT-4o	0.723	0.607	0.871
Gemini 1.5 Pro	0.755	0.652	0.893
LLaMA 3 8B	0.658	0.559	0.806
Claude 3.5	0.868	0.736	0.961
DeepSeek V3	0.810	0.697	0.903

Table 2: Model accuracy across input conditions, showing a J-curve pattern: performance drops from Standard to Local Shuffle and peaks under the Global Shuffle condition.

Figure 2: Models exhibit significant discrepancies due to contextual informativity in legal domains, with mean error rates increasing by a factor of 1.2-2× in naturalistic versus globally shuffled contexts (n=5).

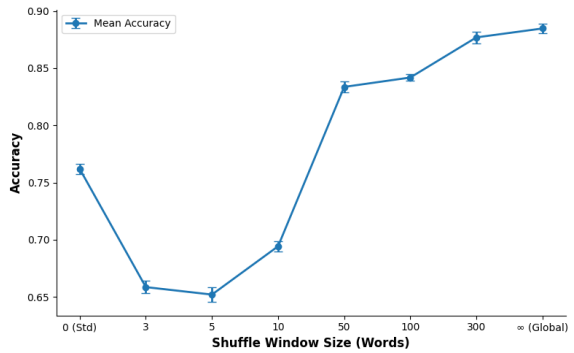

Figure 3: Mean recall across varying context shuffle sizes. Accuracy is high in standard text, drops in local context disruptions, and rises to its maximum in a global context disruptions (∞). Model: Gemini 1.5 Pro, (n=5)

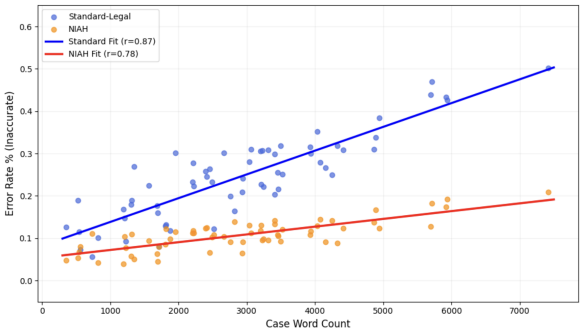

Figure 4: The performance gap between standard and NIAH tasks widens significantly with document length (p=0.0009). Model: Gemini 1.5 Pro

5 Conclusion

Reference retrieval is critical for many applications in the legal domain, for instance in determining which case texts support a particular claim. However, existing benchmarking methods do not rigorously enable evaluation of recall capabilities in previously unseen contexts. We develop an evaluation framework from U.S. court opinions that ensures models have no prior knowledge of case results or context. Applying our framework, we identify an consistent recall gap across models and tasks based on context.

We show that standard needle-in-a-haystack benchmarks consistently overestimate recall performance in the legal domain. Moreover, we isolate the causes of performance degradation to contextual informativity rather than distributional differences, and find a J-shaped curve suggesting complex reading behavior. Thus, our findings highlight the need for specialized testing in reference-critical applications, and establish an evaluation framework for improving retrieval across informativity levels.

6 Limitations

Although our methodology eliminates positional bias through ablation tests, there could be other internal mechanisms influencing accuracy within the LLMs that we have not considered. This makes it so that we cannot claim that the J-curve pattern we observe is completely responsible for the model's

behavior. Through this limitation, we open the door for future studies to probe deeper into model internals and perform extensive attention analysis to validate and explore our results further.

References

2024. Lepard: A large-scale dataset of judicial citations to precedent. In *ACL 2024 Long*.

Anthropic. 2024. Introducing claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet.

Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2024. Blt: Can large language models handle basic legal text? *Preprint*, arXiv:2311.09693.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2021. Lexglue: A benchmark dataset for legal language understanding in english. In *Findings of ACL (Long Papers) / arXiv*.

Ling Chen, Xujiang Zhao, Jiaying Lu, and et al. 2023. Domain specialization as the key to make large language models disruptive: A comprehensive survey.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models.

DeepSeek-AI. 2024. Deepseek-v3 technical report. Technical report - formal citation needed.

Yongqi Fan, Hongli Sun, Kui Xue, Xiaofan Zhang, Shaoting Zhang, and Tong Ruan. 2024. Medodyssey: A medical domain benchmark for long context evaluation up to 200k tokens. *arXiv preprint arXiv:2406.15019*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Sawada, Kwangho Tae, Anish Thite, and 3 others. 2023. A framework for few-shot language model evaluation.

Google. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Shashank Guha and 1 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning. *Preprint*, arXiv:2308.11462.

Vipul Gupta, David Pantoja, Candace Ross, Adina Williams, and Megan Ung. 2024. Changing answer order can decrease mmlu accuracy.

Ghita Houir Alami and 1 others. 2024. Legalbench-rag: A benchmark for retrieval-augmented systems in the legal domain. *Preprint*, arXiv:2408.10343.

Yinghao Hu, Leilei Gan, Wenyi Xiao, Kun Kuang, and Fei Wu. 2025. Fine-tuning large language models for improving factuality in legal question answering. *Preprint*, arXiv:2501.06521.

Zihao Li. 2023. The dark side of chatgpt: Legal and ethical challenges from stochastic parrots and hallucination. *Preprint*, arXiv:2304.14347.

Dimitris Mamakas, Petros Tsotsi, Ion Androutsopoulos, and Ilias Chalkidis. 2022. Processing long legal documents with pre-trained transformers: Modding legalbert and longformer. *Preprint*, arXiv:2211.00974.

Meta. 2024. The llama 3 herd of models. Technical report - formal citation needed.

Hyeonseok Moon and Heuiseok Lim. 2025. Needlechain: Measuring intact long-context reasoning capability of large language models. *Preprint*, arXiv:2507.22411.

OpenAI. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Yifu Qiu, Varun Embar, Yizhe Zhang, Navdeep Jaitly, Shay B. Cohen, and Benjamin Han. 2025. Eliciting in-context retrieval and reasoning for long-context large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3176–3192. Association for Computational Linguistics.

Philip Resnik. 1992. Wordnet and distributional analysis: A class-based approach to lexical discovery. In *AAAI Workshop on Statistically-Based Natural Language Processing*, AAAI Technical Report WS-92-01.

Yidong Wang. 2025. Reasoning on multiple needles in a haystack.

Yifei Yu, Qian-Wen Zhang, Lingfeng Qiao, Di Yin, Fang Li, Jie Wang, Zengxi Chen, Suncong Zheng, Xiaolong Liang, and Xing Sun. 2025. Sequential-niah: A needle-in-a-haystack benchmark for extracting sequential needles from long contexts.

Lucia Zheng and 1 others. 2025. A reasoning-focused legal retrieval benchmark. *Preprint*, arXiv:2505.03970.

A Random Chance Results for Context

To ensure that our models did not have prior context for the evaluation we conducted, we evaluated each in a no-context setting (n=5). As expected for a model with no prior knowledge, they performed at random chance.

Model	Accuracy (No Context)
GPT-4o	0.503
Gemini 1.5 Pro	0.500
LLaMA 3 8B	0.484
Claude 3.5	0.516
DeepSeek V3	0.484

Table 3: Model accuracy on legal classification task with no contextual information provided.

B Positional Ablation Tests

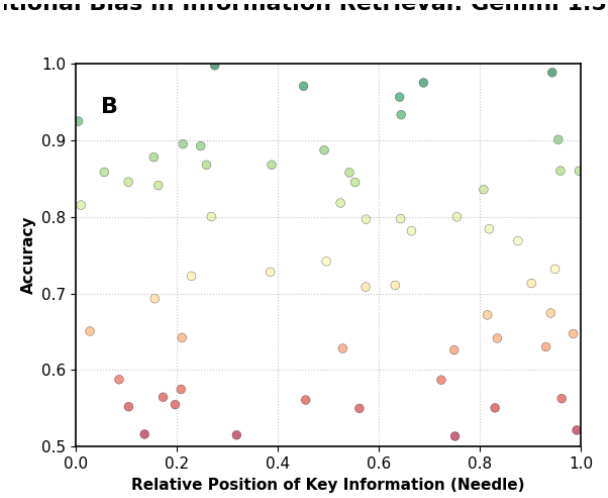

Figure 5: Our positional ablation tests show no correlation between needle position and demonstrate that position cannot explain the gap between standard and NIAH error rates.