Harnessing the Unseen: The Hidden Influence of Intrinsic Knowledge in Long-Context Language Models

Yu Fu[1], **Haz Sameen Shahgir**[1], **Hui Liu**[2], **Xianfeng Tang**[2], **Qi He**[3], **Yue Dong**[1*]
[1]University of California, Riverside,
[2]Amazon,
[3] Microsoft
yfu093@ucr.edu, yue.dong@ucr.edu

Abstract

Recent advances in long-context language models (LCLMs), designed to handle extremely long contexts, primarily focus on utilizing external contextual information, often leaving the influence of language models' parametric knowledge underexplored. In this work, we firstly investigate how this parametric knowledge affects content generation and demonstrate that its impact becomes increasingly pronounced as context length extends. Furthermore, we show that the model's ability to utilize parametric knowledge, which we call parametric recall ability, does not improve simultaneously with its ability to leverage contextual knowledge through extrinsic retrieval ability. Moreover, better extrinsic retrieval ability can interfere with the model's parametric recall ability, limiting its full potential. To bridge this gap, we design a simple yet effective Hybrid Needle-in-a-Haystack test that evaluates models based on their capabilities across both abilities, rather than solely emphasizing extrinsic retrieval ability. Our experimental results reveal that Qwen-2.5 models significantly outperform Llama-3.1 models, demonstrating superior potential to combine various abilities. Moreover, even the more powerful `Llama-3.1-70B-Instruct` model fails to exhibit better performance, highlighting the importance of evaluating models from a dual-ability perspective.

Code — https://github.com/FYYFU/Hybrid-NIAH

Introduction

Recent advancements in both open-source models (e.g., LLaMA (Grattafiori et al. 2024), Qwen (Yang et al. 2024a,b)) and closed-source models (e.g., GPT-4 (Achiam et al. 2023; OpenAI 2024), Claude (Anthropic 2024), Gemini (DeepMind 2024)) have incorporated long-context capabilities and significantly extended their context windows. For instance, GPT-4o and Claude Sonnet 4 have expanded their context windows to 128K and 200K tokens respectively. The emergence of those Long Context Language Models (LCLMs) has significantly advanced the ability of language models to process and generate coherent content over extended contexts. They have become increasingly useful for various tasks, including document summarization (Yen et al. 2025), multi-turn conversations (Maharana et al. 2024) and question answering (Karpinska et al. 2024), where retaining and utilizing long-term dependencies is essential.

With the advance of LCLMs, numerous benchmarks have been proposed to evaluate their effectiveness in handling long contexts. For example, the Needle-in-a-Haystack test (Kamradt 2023) inserts fabricated critical information (i.e. "needle") into long irrelevant documents (i.e. "haystack") to examine LCLMs' external retrieval ability and prevent the influence of their own parametric knowledge. Other benchmarks such as RULER (Hsieh et al. 2024), InfiniteBench (Zhang et al. 2024) and HELMET (Yen et al. 2025) extend to more realistic tasks such as QA, summarization etc., and ultra long contexts (100K+ tokens) to systematically assess LCLMs' capacity to process and generate meaningful outputs over extraordinarily external long contexts. These benchmarks evaluate how effectively these models leverage external long contexts, thereby emphasizing their extrinsic retrieval ability while overlooking the role of parametric knowledge (Roberts, Raffel, and Shazeer 2020; Wang, Liu, and Zhang 2021; Jiang et al. 2020).

However, the parametric knowledge has already been shown to be pivotal to the performance of language models, especially when parametric and extrinsic knowledge contradict each other (Xie et al. 2023; Wang et al. 2024a; Xu et al. 2024b). **In this work, we argue that parametric knowledge plays an important role during long-context generation, with its impact becoming more pronounced as context length extends.** To validate this hypothesis, we construct a new dataset, `IWhoQA`, based on WhoQA (Pham et al. 2024), specifically designed to systematically probe LCLMs in scenarios where parametric knowledge either aligns with or contradicts the external context. Experimental results show that LCLMs consistently perform better when the external context aligns with their parametric knowledge. The performance gap increases with longer contexts, reaching up to 10 points on `Llama-3.1-8B-Instruct`, suggesting a growing reliance on parametric knowledge during long-context generation.

Given the importance of parametric knowledge for LCLMs and the lack of its evaluation in existing benchmarks, it remains unclear whether methods that perform well in these benchmarks improve only extrinsic retrieval ability, or also enhance parametric recall ability. To address

*Corresponding Author

this gap, we systematically investigate the relationship between extrinsic retrieval and parametric recall. Specifically, we compare RoPE (Su et al. 2021) and STRING (An et al. 2024b), an improved variant of RoPE designed for long-context tasks, across various curated datasets featuring diverse knowledge alignments between external context and parametric knowledge. Our results show that while STRING improves extrinsic retrieval ability, it suppresses parametric recall even when parametric knowledge benefits generation. This reveals a trade-off between leveraging external context and utilizing parametric knowledge, an aspect overlooked by current benchmarks.

Building on these observations, we propose a simple yet effective Hybrid Needle-in-a-Haystack test to comprehensively evaluate how well models integrate parametric recall ability with extrinsic retrieval ability during long-context generation. Specifically, we design queries such as *"What's the favorite thing of the person who wrote {Book_Name}?"*—which require the model to first recall the author's name from its parametric knowledge, and then retrieve the inserted needle in the context to answer the question. Experimental results show that the Qwen2.5 family demonstrates a near-linear improvement in its ability to integrate both types of knowledge as model size increases. In contrast, the Llama3.1 family exhibits minimal gains despite a substantial increase in parameters, suggesting that it struggles to effectively leverage parametric knowledge in long-context settings. Our contributions can be summarized as follows:

- We present the first study demonstrating that the parametric knowledge of LCLMs plays an important role in long-context generation, with its influence becoming increasingly pronounced as the context length extends.
- We show that naively improving LCLMs' extrinsic retrieval ability will unintentionally suppress their parametric recall ability, revealing a trade-off between utilizing external contexts and leveraging parametric knowledge.
- We introduce a simple yet effective Hybrid Needle-in-a-Haystack test to examine LCLMs' ability to integrate the parametric recall ability and extrinsic retrieval ability during long-context generation. We show that this ability doesn't always scale with the model size, highlighting limitations in existing LCLMs.

Related Work

Parametric Knowledge During pre-training, LLMs memorize a vast amount of knowledge into their parameters, known as parametric (intrinsic) knowledge (Roberts, Raffel, and Shazeer 2020; Wang, Liu, and Zhang 2021; Jiang et al. 2020). This parametric knowledge can often conflict with new information presented via prompting through a phenomenon known as knowledge conflict (Xie et al. 2023; Longpre et al. 2022; Pan et al. 2023; Xu et al. 2024a). While earlier works have found that LLMs rely on parametric knowledge when prompted with simple counterfactual statements (Longpre et al. 2022; Chen, Zhang, and Choi 2022), more recent research has demonstrated that LLMs rely more on the information presented as the context (Xie et al. 2023; Pan et al. 2023; Wan, Wallace, and Klein 2024). Pan et al. (2023) uncovers LLMs' sensitivity to small amounts of even misinformation in context and Xu et al. (2024a) applies prompting to persuade LLMs to disregard their parametric knowledge. However, Xie et al. (2023) further finds that when presented with conflicted evidence, LLMs resolve in favor of parametric knowledge. Zhou et al. (2023) proposes framing the conflicting information as user opinion to prevent the LLM from taking it as fact. We refer the reader to Xu et al. (2024b) for further details. However, to the best of our knowledge, the relation between parametric knowledge and long-context tasks has not been studied.

Long Context Models Training transformers on long context from scratch is prohibitively expensive. Much research has focused on extending the context window of existing models through two categories: 1) Finetuning on a small amount of long context data (Peng et al. 2024; Lazarevich et al. 2025; Chen et al. 2023; jquesnelle 2023) and 2) Training-free methods based on reusing position information. (An et al. 2024a; bloc97 2023; emozilla 2023; Su 2023; Jin et al. 2024). For example, Self-Extend (Jin et al. 2024), Dual Chunk Attention (An et al. 2024a), and STRING (An et al. 2024b) all propose replacing large relative positions by overwriting them with proportionately smaller ones seen during training. By mapping large relative positions to smaller ones that are well represented during pre-training, the model can more effectively integrate provided information into long-context generation.

Long Context Evaluation Multiple synthetic tasks have been proposed to evaluate LCLMs, including Needle-in-a-Haystack (NIAH) test (Kamradt 2023), long-context QA, summarization, reasoning, extrinsic learning (ICL) and retrieval-augmented generation (RAG), among others (Hsieh et al. 2024; Kim et al. 2025; Gao et al. 2025; Liu et al. 2024; Yen et al. 2025). Variants of NIAH include multiple-needle, long-needle, and needle-in-needle configurations (Gao et al. 2025) and have also been extended to test long-context reasoning (Kuratov et al. 2024), mathematical (Wang et al. 2024b) and multilingual capabilities (Hengle et al. 2024). However, several works have found that the standard NIAH task may not truly reflect an LCLM's downstream capabilities and have proposed comprehensive benchmarks such as RULER (Hsieh et al. 2024) and its multilingual extension ONERULER (Kim et al. 2025), LongBench (Bai et al. 2024) , LongGenBench (Liu et al. 2024), L-Eval (An et al. 2023), InfiniteBench (Zhang et al. 2024) and HELMET (Yen et al. 2025). These benchmarks contain several tasks such as citation generation, reranking, QA, summarization, ICL, and RAG. Karpinska et al. (2024) introduced NoCha, a dataset of human-written true and false claims about recently published fictional books. While previous work has primarily studied long context in isolation, some going so far as to use entirely fictional contexts (Gao et al. 2025; Karpinska et al. 2024), to the best of our knowledge the problem of how an LCLM's parametric knowledge affects its long context abilities has not been explored.

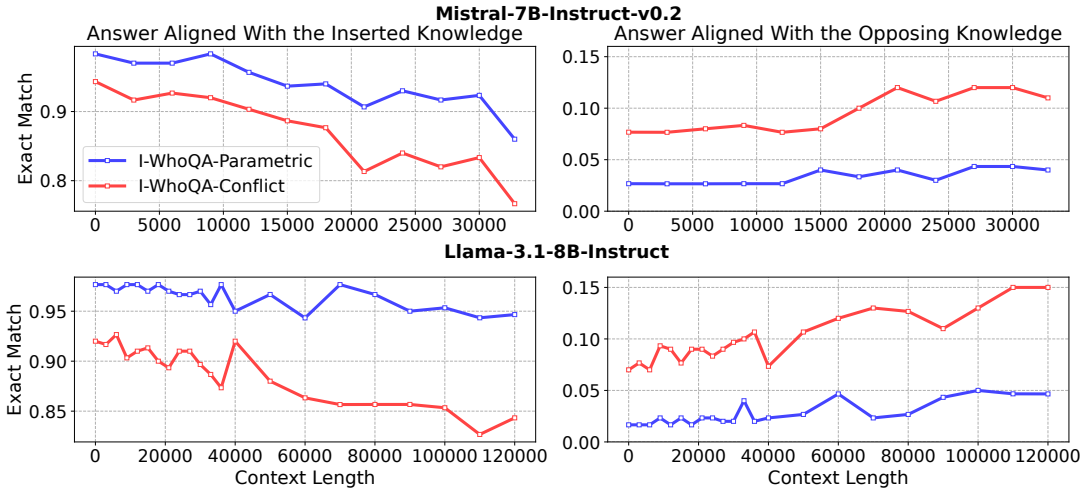

Figure 1: We only count answers that align with either the injected knowledge (whether parametric or conflictual) or the opposing knowledge source. Left) LCLMs struggle to retrieve the answer from the context on the I-WhoQA-Conflict subset, i.e., when the context information conflicts with their parametric knowledge. Right) The upward trend of I-WhoQA-conflict (red) shows that when the parametric knowledge conflicts with the context, the likelihood of a LCLM relying on parametric knowledge steadily increases with larger contexts.

The Role of Parametric Knowledge in Long-Context Generation

Existing benchmarks for evaluating LCLMs primarily focus on assessing their extrinsic retrieval ability—how well they extract information or perform reasoning based solely on the external context. In contrast, our work aims to investigate the role of parametric knowledge embedded within LCLMs. We begin by posing a central question: **What role does the parametric knowledge of LCLMs play in long-context generation?** Specifically, we explore how parametric knowledge interacts with external context and how this interplay affects the quality of the generated output.

Dataset Creation

While existing datasets in the knowledge conflict (Xie et al. 2023; Xu et al. 2024b; Pham et al. 2024) also aim to examine the role of parametric knowledge in LLMs, they are not designed for evaluating LCLMs. To address this gap, we first construct a dataset to investigate whether LCLMs naturally incorporate parametric knowledge during long-context generation across varying context lengths. Specifically, we choose WhoQA (Pham et al. 2024), a short-form QA dataset that provides multiple context-answer pairs for each entity, as the foundation for our construction.

Specifically, for each entity in WhoQA, we generated an answer to every associated question in the dataset and retained only those entities where the model consistently produced a single, invariant answer. This filtering process ensures that the model's parametric knowledge encodes a single, unambiguous answer for the retained entities. Since parametric knowledge may differ across LCLMs, we constructed a separate dataset of 300 examples for each model, which we refer to as the **I-WhoQA** dataset.

To assess how parametric knowledge influences long-context generation, we select two distinct context-answer pairs for each example in the I-WhoQA dataset: (1) a context that aligns with the model's parametric knowledge, referred to as the **I-WhoQA-Parametric** subset; (2) a context that conflicts with the model's parametric knowledge, referred to as the **I-WhoQA-Conflict** subset. By comparing model performance across these two subsets and varying context lengths, we aim to reveal the extent to which parametric knowledge impacts long-context generation. An example from the I-WhoQA dataset is provided in the Appendix.

Experiment

Experimental Setting To validate the role of parametric knowledge in long-context generation, we conducted experiments using two curated I-WhoQA subsets mentioned above, alongside two different backend models — `mistral-7b-instruct-v0.2` (Jiang et al. 2023) and `Llama3.1-8B-instruct` (Grattafiori et al. 2024). For each backend model, we employed greedy decoding strategy with a maximum of 32 generation tokens. To investigate the effect of context length, a key factor in long-context generation, we report results across varying context lengths. For each example, the only difference across these different context settings is the amount of irrelevant content inserted into the context, which is consistent with the setup of Needle-in-a-Haystack test. In our evaluation, we only include answers that clearly align with one of two competing knowledge sources: the injected context (which can represent either parametric or conflict knowledge) or the alternative source (i.e., the one not injected). This ensures a clean comparison between model behavior in context-following vs. parametric recall scenarios.

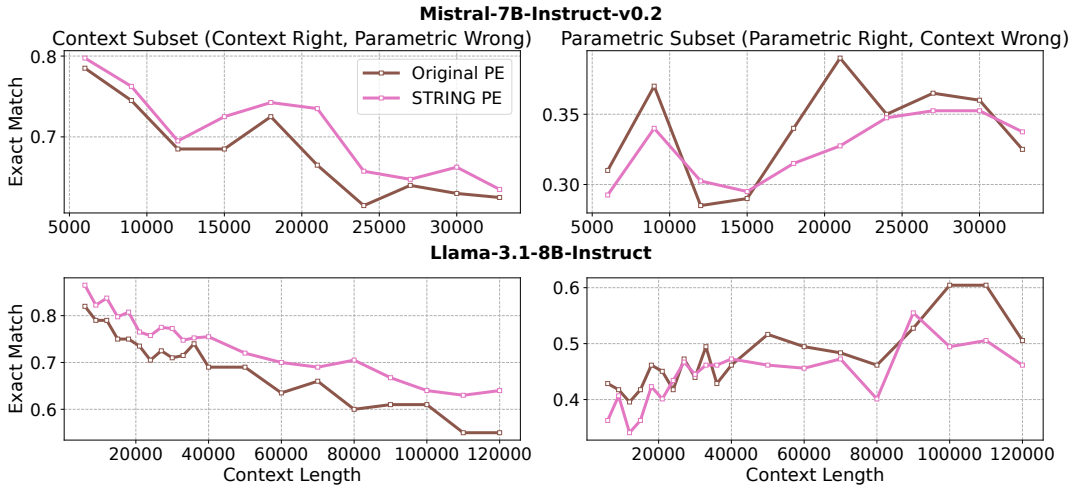

Figure 2: Performance of LCLMs on HotPotQA-Context and -Parametric subsets. Left) STRING improves performance on the HotPotQA-context subset by enhancing extrinsic retrieval ability. Right) STRING hinders the model's ability to recall parametric knowledge and leads to the decrease of performance.

Main Results Results in Figure 1 show that the parametric knowledge of LCLMs influences long-context generation and this influence becomes more pronounced as context length increases. First, the left side of Figure 1 (Answer Aligned With the Inserted Knowledge) shows the performance on the IWhoQA-Conflict subset is consistently worse than on IWhoQA-Parametric subset. This demonstrates that LCLMs are struggling to follow the given context when its parametric knowledge contains a different answer. Second, from the right side of Figure 1 (Answer Aligned With the Opposing Knowledge) on the I-WhoQA-Conflict subset, where conflicting contexts are provided as supporting facts, we observed a notable increase in the proportion of outputs adhering to opposing knowledge (i. e. parametric knowledge) as context length increased. This trend is especially pronounced when `Llama-3.1-8B-Instruct` serves as the backend model, which can handle longer context lengths. **This suggests that as the context length increases, the model increasingly relies on its parametric knowledge even when that knowledge contradicts the external context.**

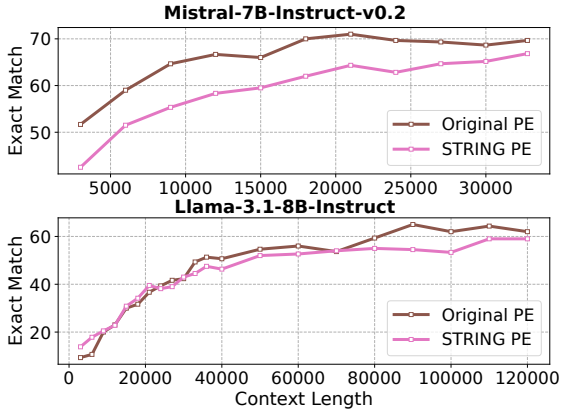

Figure 3: I-WhoQA-Irrelevant subsets: As the length of the context is increased, LCLMs ignore the context and generate according to their parametric knowledge. However, using STRING causes LCLMs to re-focus on irrelevant context and generate wrong answers.

The Trade-off Between Parametric Recall Ability and Extrinsic Retrieval Ability

In Section 1, we demonstrated that the parametric knowledge within LCLMs increasingly influences generation as context length grows. While existing benchmarks and research (An et al. 2024b; Han et al. 2024; emozilla 2023) on LCLMs have largely focused on evaluating and improving extrinsic retrieval, little attention has been paid to how parametric knowledge affects long-context generation. In this section, we aim to answer the question: **Can we improve extrinsic retrieval without compromising parametric recall in long-context generation?** We examine whether STRING (An et al. 2024b), an improved variant of RoPE designed for long-context modeling, can enhance both capabilities. Experimental results reveal that although STRING improves extrinsic retrieval, it can inadvertently impair parametric recall, sometimes leading to a decline in overall model performance.

Dataset Creation

In Section 1, we constructed the I-WhoQA dataset, ensuring that all examples can be answered using the model's parametric knowledge. To further assess the trade-off between extrinsic retrieval and parametric recall, we created the **I-WhoQA-Irrelevant** subset. In this subset, instead of providing context relevant to the current question, we inserted entirely irrelevant content to evaluate how models handle unrelated information, and whether STRING remains effective

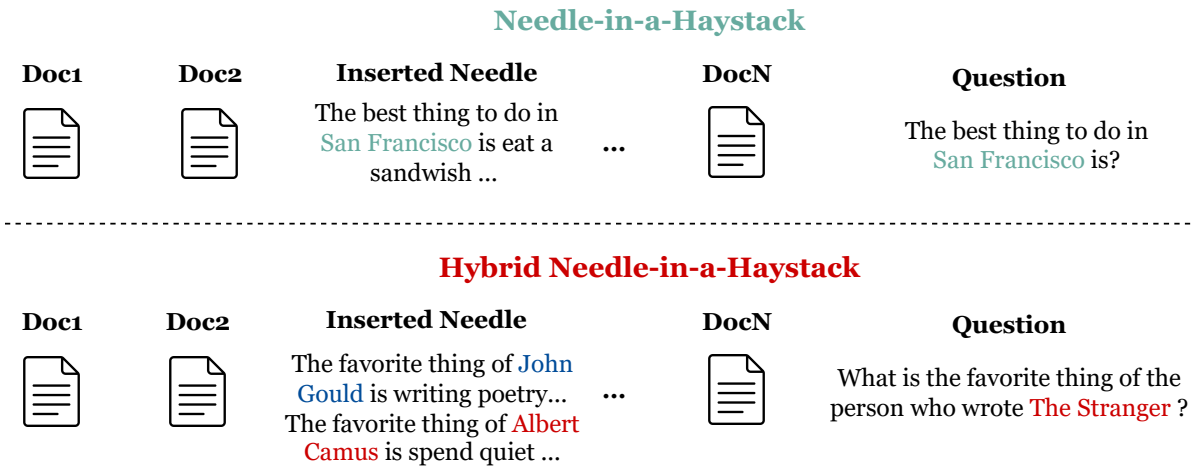

Figure 4: Upper) Needle-in-a-Haystack. It involves directly inserting the answer into the haystack and retrieving it. Lower) Hybrid Needle-in-a-Haystack. It requires a two-step process: first, performing parametric recall to identify the retrieval target based on the model's parametric knowledge, and then retrieving the answer from the haystack.

when answering the question requires parametric knowledge rather than context-based extrinsic retrieval.

In addition to the **I-WhoQA-Irrelevant** subset, we construct a more realistic dataset based on HotpotQA (Yang et al. 2018) to simulate scenarios where the context appears relevant to the question but ultimately fails to support the correct answer. From the HotpotQA dataset, we derive two subsets by analyzing the relationship between the external context, the model's parametric knowledge, and the reference answer.

First, to distinguish the role of external context from that of the model's parametric knowledge, we ensured that the answers derived from these two sources are not the same. Building on this distinction, we obtained two final subsets. The first subset, referred to as the **HotpotQA-Context** subset, contains examples where the reference answer can be derived from the given contexts rather than the parametric knowledge. The second subset, known as the **HotpotQA-Parametric** subset, consists of examples where the reference answer matches the model's parametric knowledge. We constructed a dataset consisting of 200 examples for each subset, except for the HotpotQA-Parametric subset for `Llama3.1-8B-Instruct`, which contains a total of 93 examples.

Experiment

Experimental Setting We use the same backend models and generation configuration as described in Section 1, applied to the newly constructed datasets. For STRING, we configure the local value to 128 and two different shift-ratios: 0.25 and 0.33. We report the average performance across these two settings. More details about STRING can be found in An et al. (2024b).

Main Results As shown in Figure 3, STRING consistently underperforms the baseline RoPE method in terms of accuracy. This performance degradation can be attributed to STRING's stronger extrinsic retrieval bias in long-context settings, which causes LCLMs to incorrectly focus on irrelevant context. Even when the external context is entirely unrelated to the question, STRING tends to extract the answer from the context rather than leveraging the model's parametric knowledge, which actually contains the correct answer.

Results on HotpotQA-Context subset and HotpotQA-Parametric subset are shown in Figure 2. STRING significantly improves performance on the HotpotQA-Context subset, where the model must rely on the external context to derive the correct answer. This confirmed that STRING is improving extrinsic retrieval ability on LCLMs. However, on the HotpotQA-Parametric subset where the context is related but not useful, STRING underperforms RoPE and this is consistent with our finding on the I-WhoQA-Irrelevant subset. **In conclusion, STRING's enhanced extrinsic retrieval ability interferes with the model's use of parametric knowledge, leading to degraded performance in scenarios where the context is misleading or insufficient.** This trade-off becomes more pronounced as the context length increases since STRING's boosted extrinsic retrieval ability can be more prominent with longer contexts as shown on the `Llama-3.1-8B-Instruct` HotPotQA-Context subset.

Hybrid Needle-in-a-Haystack Framework

So far, we have shown that parametric knowledge plays a crucial role in the long-context generation, yet improvements in extrinsic retrieval do not translate into better parametric recall. Moreover, existing benchmarks (Hsieh et al. 2024; An et al. 2023; Zhang et al. 2024) primarily focus on how to utilize external context, overlooking scenarios that require interaction with parametric knowledge. As a result, methods (An et al. 2024b) developed under these

Model	Generation Length = 32				Generation Length = 64			
	Random Facts				Random Facts			
	0	1	2	3	0	1	2	3
Needle-in-a-Haystack								
Mistral-7B-Instruct-v0.2	100	100	99.55	98.41	-	-	-	-
Llama-3.1-8B-Instruct	100	99.21	99.33	98.83	-	-	-	-
Qwen2.5-7B-Instruct	99.78	99.64	99.81	99.92	-	-	-	-
Hybrid Needle-in-a-Haystack								
Mistral-7B-Instruct-v0.2	89.31	58.30	53.53	53.21	97.83	63.28	58.16	58.11
Mistral-7B-Instruct-v0.3	73.17	59.54	56.97	56.94	76.56	62.48	58.01	57.73
Llama-3.1-8B-Instruct	92.97	83.66	72.60	67.47	96.71	90.28	80.93	74.39
Llama-3.1-70B-Instruct	95.73	86.39	74.48	71.81	95.77	89.88	77.66	74.68
Qwen2.5-7B-Instruct	86.38	83.32	77.23	73.14	90.72	88.46	80.98	77.03
Qwen2.5-14B-Instruct	54.46	76.74	75.32	73.97	92.95	94.32	89.85	85.60
Qwen2.5-72B-Instruct	64.01	98.13	97.92	97.77	94.16	99.53	99.84	98.86
Qwen2.5-7B-Instruct-1M-128k	97.72	84.83	82.10	79.31	98.50	84.89	82.22	79.37
Qwen2.5-7B-Instruct-1M	60.07	48.09	43.70	42.91	60.77	49.26	43.43	43.21

Table 1: Average accuracy score of different models. `Qwen2.5-7B-Instruct-1M-128k` means we set the max input length to 128k.

benchmarks fail to achieve better performance when parametric recall is essential. **To address this gap, we propose a simple yet effective Hybrid Needle-in-a-Haystack (NIAH) test designed to jointly evaluate parametric recall ability and extrinsic retrieval ability.** This test enables a more comprehensive assessment of how these two forms of knowledge interact, complement, or interfere with one another, ultimately affecting the quality of model generation.

Dataset Creation

NIAH test uses irrelevant content as a haystack and inserts the answer from a question-answer pair into the haystack to evaluate LCLMs' extrinsic retrieval ability under different context lengths and insertion depths. Unlike the standard NIAH test, our Hybrid NIAH test simultaneously assesses parametric recall ability and extrinsic retrieval ability by designing scenarios where the model must first utilize its parametric knowledge and then perform extrinsic retrieval on the haystack. A comparison between NIAH test and Hybrid NIAH test is illustrated in Figure 4. The Hybrid NIAH test introduces an additional step designed to trigger the parametric recall ability, requiring the model to recall the relevant target from parametric knowledge. The model relies on this recalled target entity to guide extrinsic retrieval from the external context, ultimately producing the final answer.

Firstly, to implement the parametric recall step, we use the I-WhoQA dataset constructed in Section 1. To ensure that the question-answer pairs only require knowledge that exists across all models, we perform an intersection operation between the I-WhoQA datasets obtained from `Mistral-7B-Instruct-v0.2` and `Llama3.1-8B-Instruct`. Secondly, for the extrinsic retrieval step, we adopt a similar approach to the NIAH test and create various random facts to serve as inserted answers to the corresponding questions. For the remaining irrelevant haystack construction, we follow the NIAH test and use the PaulGrahamEssays dataset to construct the target length haystack.

To prevent the model from exploiting potential syntactic patterns present in the question-answer pairs to extract the inserted "needle" from the context, we introduce varying amounts of random facts as interference (from 0 to 3). This additional noise ensures that models understand the task and the context rather than relying on superficial patterns, thereby enhancing the robustness and reliability of the final result. Details of the final knowledge pairs are listed in the Appendix.

Experiment

Experimental Setting We evaluated models from three different families with various specifications: Mistral, Llama3.1, and Qwen2.5. For all models, we utilized the Instruct versions and employed greedy decoding strategy. We set the maximum generation length to 32 and 64. For each model, we divided the maximum context length into 40 intervals. For each interval, we defined 10 progressively increasing insertion depths and conducted Hybrid NIAH tests accordingly. This setup allows us to systematically evaluate the models' performance across varying context lengths and insertion depths. To ensure the stability of the results, we use multiple examples and report the average result. Finally, to measure the similarity between the predictions and the references, we adopt the method provided by PyramidKV (Cai et al. 2024), which calculates the final result based on exact matches between tokens. The formula is as follows:

$$Score = \frac{|P| \cap |A|}{|A|}$$

where P and A are tokens from predictions and references respectively. More detail settings are listed in the Appendix.

Main Results The averaged results for the Hybrid NIAH test are presented in Table 1 and we list the detailed results in the Appendix. From these results, we can observe three interesting findings.

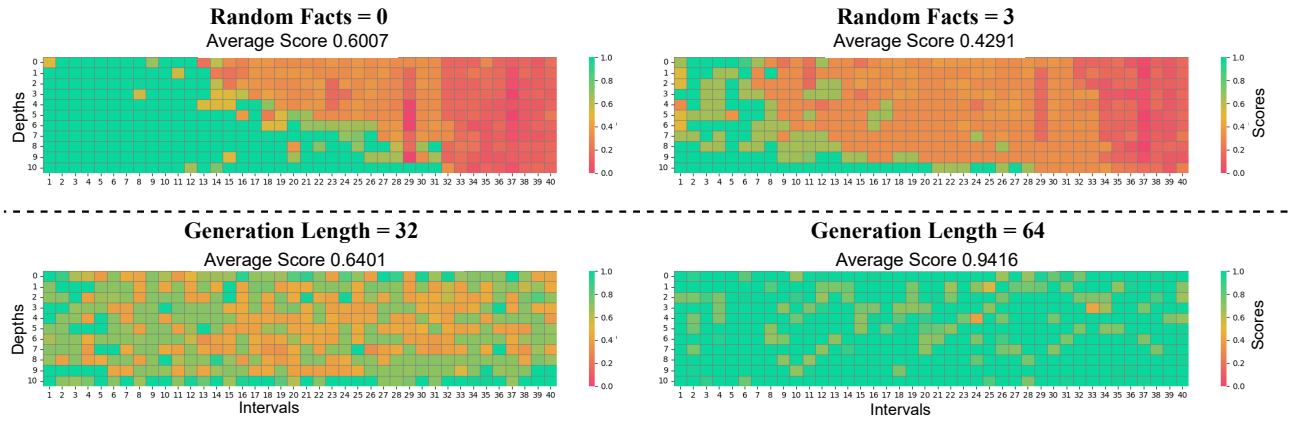

Figure 5: Hybrid NIAH test results. Upper) `Qwen2.5-7B-Instruct-1M` with generation length 32 and various numbers of inserted random facts. Lower) `Qwen2.5-72B-Instruct` with generation lengths 32 and 64. The number of random facts was set to 0.

Firstly, with only a simple modification to the question, our Hybrid NIAH test becomes significantly more challenging than the standard NIAH test, as shown in Table 1. This effect is especially evident in long-context scenarios, as demonstrated by the results on `Qwen2.5-7B-Instruct-1M` in Figure 5. Moreover, by inserting varying amounts of random facts, we observe a clear decline in model performance by up to 25% for `Llama-3.1-70B-Instruct`, while the NIAH test results remain unaffected by this interference, as shown in Table 1. This discrepancy indicates that models rely on superficial patterns within the context to extract answers when there is a single needle, rather than genuinely leveraging their parametric knowledge. The insertion of random facts effectively disrupts this pattern-based retrieval process, allowing us to better assess whether a model truly utilizes its parametric knowledge rather than merely exploiting contextual shortcuts. Multiple-Needles Hybrid NIAH test serves as a more rigorous evaluation of how well models integrate their parametric recall ability with the long-context generation.

Secondly, models from the Mistral and Llama3.1 families fail to achieve significant improvements on the Hybrid-NIAH test, even when upgraded to larger versions. This suggests that these models struggle to effectively utilize their parametric knowledge under long-context scenarios. Despite increased model capacity, their performance remains constrained when external context is incomplete or ambiguous. In contrast, the Qwen2.5 family shows substantial improvements as model size scales up. Larger Qwen2.5 models are more capable of effectively leveraging their parametric knowledge, providing more accurate and factually consistent results even when the external context is incomplete.

Finally, we observe an interesting phenomenon in the Qwen2.5 models for Single-Needle Hybrid NIAH test: larger model variants initially claim to not see the inserted needle but are able to generate and then retrieve it eventually. As shown in the Appendix, `Qwen2.5-14B-Instruct` and `Qwen2.5-72B-Instruct` start by generating a refusal but then successfully retrieve the needle regardless. In contrast, the smaller `Qwen2.5-7B-Instruct` directly retrieves the needle without any refusal. This behavior explains the findings in Table 1 where `Qwen2.5-14B-Instruct` and `Qwen2.5-72B-Instruct` underperforming `Qwen2.5-7B-Instruct` when generation length is 32. Limiting the generation length only allows the larger Qwen models to generate the initial refusal and not the retrieved needle. Strangely, this refusal behavior is not observed when there are multiple inserted needles. Since this behavior is unique to the Qwen2.5 family and only for single Hybrid NIAH test, we hypothesize it is related to the Qwen's chunked attention mechanism (An et al. 2024a) struggling to effectively focus on the singular inserted needle when generating the first token. When multiple needles are inserted, the syntactical similarity between the needles helps the larger Qwen models to focus on the context and they do not refuse to answer initially. We leave further investigation of this phenomenon for future work.

Conclusion

In this work, we investigate the role of parametric knowledge in LCLMs and uncover a critical trade-off between parametric recall and extrinsic retrieval. Our analysis reveals that enhancements achieved on extrinsic retrieval ability can inadvertently suppress the model's use of its parametric knowledge, especially when the context is irrelevant or misleading. To evaluate this interplay, we introduce the novel Hybrid Needle-in-a-Haystack test that assesses a model's ability to jointly integrate parametric and extrinsic knowledge. Experimental results demonstrate that even large-scale LCLMs struggle to effectively combine these two abilities and often fail to fully leverage their parametric knowledge during long-context generation. Our findings highlight the need for future model designs and evaluation protocols that consider both parametric and extrinsic knowledge sources, paving the way for more robust and context-aware LCLMs.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

An, C.; Gong, S.; Zhong, M.; Li, M.; Zhang, J.; Kong, L.; and Qiu, X. 2023. L-Eval: Instituting Standardized Evaluation for Long Context Language Models. arXiv:2307.11088.

An, C.; Huang, F.; Zhang, J.; Gong, S.; Qiu, X.; Zhou, C.; and Kong, L. 2024a. Training-free long-context scaling of large language models. *arXiv preprint arXiv:2402.17463*.

An, C.; Zhang, J.; Zhong, M.; Li, L.; Gong, S.; Luo, Y.; Xu, J.; and Kong, L. 2024b. Why Does the Effective Context Length of LLMs Fall Short? arXiv:2410.18745.

Anthropic. 2024. Introducing the next generation of Claude. https://www.anthropic.com/news/claude-3-family.

Bai, Y.; Lv, X.; Zhang, J.; Lyu, H.; Tang, J.; Huang, Z.; Du, Z.; Liu, X.; Zeng, A.; Hou, L.; Dong, Y.; Tang, J.; and Li, J. 2024. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding. arXiv:2308.14508.

bloc97. 2023. NTK-Aware Scaled RoPE Allows LLaMA Models to Have Longer Contexts. Accessed: 2025-03-22.

Cai, Z.; Zhang, Y.; Gao, B.; Liu, Y.; Liu, T.; Lu, K.; Xiong, W.; Dong, Y.; Chang, B.; Hu, J.; et al. 2024. Pyramidkv: Dynamic kv cache compression based on pyramidal information funneling. *arXiv preprint arXiv:2406.02069*.

Chen, H.-T.; Zhang, M.; and Choi, E. 2022. Rich Knowledge Sources Bring Complex Knowledge Conflicts: Recalibrating Models to Reflect Conflicting Evidence. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2292–2307. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Chen, S.; Wong, S.; Chen, L.; and Tian, Y. 2023. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

DeepMind, G. 2024. Introducing Gemini 2.0: our new AI model for the agentic era. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/#project-mariner.

emozilla. 2023. Dynamically Scaled RoPE Further Increases Context Length in LLaMA Models. Accessed: 2025-03-22.

Gao, Y.; Xiong, Y.; Wu, W.; Huang, Z.; Li, B.; and Wang, H. 2025. U-NIAH: Unified RAG and LLM Evaluation for Long Context Needle-In-A-Haystack. *arXiv preprint arXiv:2503.00353*.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Han, C.; Wang, Q.; Peng, H.; Xiong, W.; Chen, Y.; Ji, H.; and Wang, S. 2024. LM-Infinite: Zero-Shot Extreme Length Generalization for Large Language Models. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 3991–4008. Mexico City, Mexico: Association for Computational Linguistics.

Hengle, A.; Bajpai, P.; Dan, S.; and Chakraborty, T. 2024. Multilingual Needle in a Haystack: Investigating Long-Context Behavior of Multilingual Large Language Models. arXiv:2408.10151.

Hsieh, C.-P.; Sun, S.; Kriman, S.; Acharya, S.; Rekesh, D.; Jia, F.; Zhang, Y.; and Ginsburg, B. 2024. RULER: What's the Real Context Size of Your Long-Context Language Models? arXiv:2404.06654.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de Las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. *ArXiv*, abs/2310.06825.

Jiang, Z.; Araki, J.; Ding, H.; and Neubig, G. 2020. How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering. *Transactions of the Association for Computational Linguistics*, 9: 962–977.

Jin, H.; Han, X.; Yang, J.; Jiang, Z.; Liu, Z.; Chang, C.-Y.; Chen, H.; and Hu, X. 2024. LLM Maybe LongLM: Self-Extend LLM Context Window Without Tuning. arXiv:2401.01325.

jquesnelle. 2023. NTK-Aware interpolation "by parts". GitHub Pull Request, Accessed: 2025-03-22.

Kamradt, G. 2023. Needle In A HayStack. https://github.com/gkamradt/LLMTest_NeedleInAHaystack.

Karpinska, M.; Thai, K.; Lo, K.; Goyal, T.; and Iyyer, M. 2024. One Thousand and One Pairs: A "novel" challenge for long-context language models. arXiv:2406.16264.

Kim, Y.; Russell, J.; Karpinska, M.; and Iyyer, M. 2025. One ruler to measure them all: Benchmarking multilingual long-context language models. *arXiv preprint arXiv:2503.01996*.

Kuratov, Y.; Bulatov, A.; Anokhin, P.; Rodkin, I.; Sorokin, D.; Sorokin, A.; and Burtsev, M. 2024. BABILong: Testing the Limits of LLMs with Long Context Reasoning-in-a-Haystack. arXiv:2406.10149.

Lazarevich, I.; Bick, D.; Gupta, H.; Mukherjee, S.; Neema, N.; Ramakrishnan, G.; and Venkatesh, G. 2025. Extending LLM Context with 99% Less Training Tokens. https://cerebras.ai/blog/extending-llm-context-with-99-less-training-tokens.

Liu, X.; Dong, P.; Hu, X.; and Chu, X. 2024. LongGenBench: Long-context Generation Benchmark. arXiv:2410.04199.

Longpre, S.; Perisetla, K.; Chen, A.; Ramesh, N.; DuBois, C.; and Singh, S. 2022. Entity-Based Knowledge Conflicts in Question Answering. arXiv:2109.05052.

Maharana, A.; Lee, D.-H.; Tulyakov, S.; Bansal, M.; Barbieri, F.; and Fang, Y. 2024. Evaluating very long-term conversational memory of llm agents. *arXiv preprint arXiv:2402.17753*.

OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/.

Pan, L.; Chen, W.; Kan, M.-Y.; and Wang, W. Y. 2023. Attacking Open-domain Question Answering by Injecting Misinformation. In *International Joint Conference on Natural Language Processing and Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL)*, 525–539. Nusa Dua, Bali: Association for Computational Linguistics.

Peng, B.; Quesnelle, J.; Fan, H.; and Shippole, E. 2024. YaRN: Efficient Context Window Extension of Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Pham, Q. H.; Ngo, H.; Luu, A. T.; and Nguyen, D. Q. 2024. Who's Who: Large Language Models Meet Knowledge Conflicts in Practice. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Roberts, A.; Raffel, C.; and Shazeer, N. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model? In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5418–5426. Online: Association for Computational Linguistics.

Su, J. 2023. Rectified Rotary Position Embeddings. https://github.com/bojone/rerope.

Su, J.; Lu, Y.; Pan, S.; Wen, B.; and Liu, Y. 2021. RoFormer: Enhanced Transformer with Rotary Position Embedding. *ArXiv*, abs/2104.09864.

Wan, A.; Wallace, E.; and Klein, D. 2024. What Evidence Do Language Models Find Convincing? In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7468–7484. Bangkok, Thailand: Association for Computational Linguistics.

Wang, C.; Liu, P.; and Zhang, Y. 2021. Can Generative Pre-trained Language Models Serve As Knowledge Bases for Closed-book QA? *ArXiv*, abs/2106.01561.

Wang, F.; Wan, X.; Sun, R.; Chen, J.; and Arık, S. 2024a. Astute RAG: Overcoming Imperfect Retrieval Augmentation and Knowledge Conflicts for Large Language Models. arXiv:2410.07176.

Wang, L.; Dong, S.; Xu, Y.; Dong, H.; Wang, Y.; Saha, A.; Lim, E.-P.; Xiong, C.; and Sahoo, D. 2024b. MathHay: An Automated Benchmark for Long-Context Mathematical Reasoning in LLMs. arXiv:2410.04698.

Xie, J.; Zhang, K.; Chen, J.; Lou, R.; and Su, Y. 2023. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Xu, R.; Lin, B.; Yang, S.; Zhang, T.; Shi, W.; Zhang, T.; Fang, Z.; Xu, W.; and Qiu, H. 2024a. The Earth is Flat because...: Investigating LLMs' Belief towards Misinformation via Persuasive Conversation. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16259–16303. Bangkok, Thailand: Association for Computational Linguistics.

Xu, R.; Qi, Z.; Wang, C.; Wang, H.; Zhang, Y.; and Xu, W. 2024b. Knowledge Conflicts for LLMs: A Survey. *arXiv preprint arXiv:2403.08319*.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; Dong, G.; Wei, H.; Lin, H.; Tang, J.; Wang, J.; Yang, J.; Tu, J.; Zhang, J.; Ma, J.; Xu, J.; Zhou, J.; Bai, J.; He, J.; Lin, J.; Dang, K.; Lu, K.; Chen, K.; Yang, K.; Li, M.; Xue, M.; Ni, N.; Zhang, P.; Wang, P.; Peng, R.; Men, R.; Gao, R.; Lin, R.; Wang, S.; Bai, S.; Tan, S.; Zhu, T.; Li, T.; Liu, T.; Ge, W.; Deng, X.; Zhou, X.; Ren, X.; Zhang, X.; Wei, X.; Ren, X.; Fan, Y.; Yao, Y.; Zhang, Y.; Wan, Y.; Chu, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Fan, Z. 2024a. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2024b. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W. W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yen, H.; Gao, T.; Hou, M.; Ding, K.; Fleischer, D.; Izsak, P.; Wasserblat, M.; and Chen, D. 2025. HELMET: How to Evaluate Long-Context Language Models Effectively and Thoroughly. In *International Conference on Learning Representations (ICLR)*.

Zhang, X.; Chen, Y.; Hu, S.; Xu, Z.; Chen, J.; Hao, M. K.; Han, X.; Thai, Z. L.; Wang, S.; Liu, Z.; and Sun, M. 2024. ∞Bench: Extending Long Context Evaluation Beyond 100K Tokens. arXiv:2402.13718.

Zhou, W.; Zhang, S.; Poon, H.; and Chen, M. 2023. Context-faithful Prompting for Large Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 14544–14556. Singapore: Association for Computational Linguistics.