L-CiteEval: A Suite for Evaluating Fidelity of Long-context Models

Anonymous ACL submission

Abstract

Long-context models (LCMs) have witnessed remarkable advancements in recent years, facilitating real-world tasks like long-document QA. The success of LCMs is founded on the hypothesis that the model demonstrates strong **fidelity**, enabling it to respond based on the provided long context rather than relying solely on the intrinsic knowledge acquired during pre-training. Yet, in this paper, we find that open-sourced LCMs are not as faithful as expected. We introduce *L-CiteEval*, an out-of-the-box suite that can assess both generation quality and fidelity in long-context understanding tasks. It covers 11 tasks with context lengths ranging from 8K to 48K and a corresponding automatic evaluation pipeline. Evaluation of 11 cutting-edge closed-source and open-source LCMs indicates that, while there are minor differences in their generation, open-source models significantly lag behind closed-source counterparts in terms of fidelity. Furthermore, we analyze the benefits of citation generation for LCMs from both the perspective of explicit model output and the internal attention mechanism[1].

1 Introduction

The appealing long-context processing capabilities benefit large language models (LLMs) in numerous aspects (Mosbach et al., 2023; Bertsch et al., 2024), addressing areas that were once the model's blind spots, such as 1) dynamic knowledge, and 2) compatibility with efficient methodologies, such as Retrieval-Augmented Generation (RAG) (Verma, 2024). The above success stems from a strong assumption that long-context models (LCMs) possess a strong **fidelity** (Manna and Sett, 2024), which allows the models to *respond based on the given context rather than relying solely on the intrinsic knowledge acquired during pre-training*.

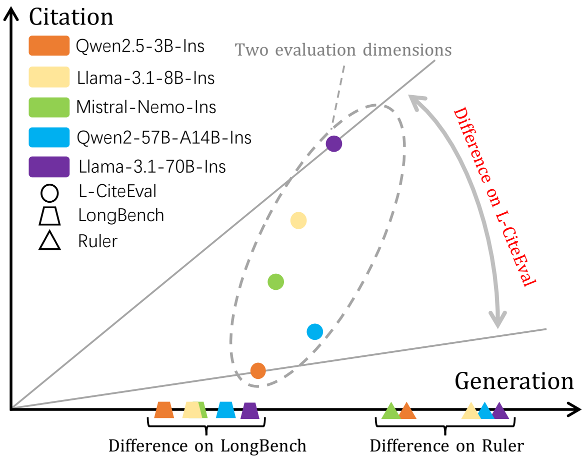


Figure 1: Comparison between *L-CiteEval* and other commonly-used long-context benchmarks, where our method evaluates LCMs from two distinct dimensions, i.e., citation quality and generation quality, amplifying the performance differences between LCMs.

Currently, most benchmarks for LCMs evaluate the model's performance by measuring the similarity of its generation to the ground truth (An et al., 2023; Li et al., 2023b; Zhang et al., 2024a). However, this results in limited differentiation among LCMs in terms of their generation capabilities. Furthermore, this paper reveals that *while the overall quality of responses across different models appears similar, their adherence to the provided context varies significantly*. This discrepancy arises because LCM performance can be affected by dataset shortcuts (Yang et al., 2024b) or potential test data leakage (Ni et al., 2024), leading to an unfair and potentially misleading evaluation. Thus, even when LCMs perform well on specific benchmarks, they may fail to generalize effectively to other tasks.

To mitigate the above issues in long-context evaluation field, we propose an out-of-the-box evaluation suite, *L-CiteEval*, which requires LCMs to generate both the statements and their supporting evidence (citations). This suite comprises two key

---

[1] Code and data are available at `https://anonymous.4open.science/r/L-CiteEval-Ana-46BF`

components: (1) a comprehensive benchmark encompassing **5** major task categories and **11** diverse long-context tasks, with context lengths ranging from **8K** to **48K**; and (2) corresponding automatic evaluation metrics and verification pipelines to ensure robust and reliable assessment. Furthermore, to disentangle the effects of task difficulty and content length, we design two controlled testing sets based on L-CiteEval: L-CiteEval-Length and L-CiteEval-Hardness. To ensure the benchmark quality, we introduce two crucial steps during the construction process: (1) we incorporate four of the latest long-context tasks into L-CiteEval, to address the challenges of timeliness and the risk of data leakage during testing (Ni et al., 2024; Apicella et al., 2024); and (2) during the dataset length expansion process, we design a rigorous padding method to avoid the impact of padding context on the model prediction.

We evaluate 11 cutting-edge and widely-used LCMs, including 3 closed-source models and 8 open-source models, with varying sizes and architectures. As shown in Fig. 1, by evaluating with L-CiteEval, the differences between LCMs become larger compared to the differences on other commonly used benchmarks. In summary, we observe that open-source models tend to rely more heavily on their intrinsic knowledge rather than on the provided context. This behavior may lead to the performance bottleneck observed in open-source LCMs (Hsieh et al., 2024). We also investigate commonly used methods in the long-context domain, including inference efficiency (Xiao et al., 2024c) and context compression methods (Verma, 2024), which reveal a trade-off between generation (efficiency or performance) and citation. In addition, we take both the model's implicit information retrieval (Wu et al., 2024) and explicit citation processes into consideration and reveal a correlation between these two manners.

2 Related Works

2.1 Long-context Understanding Benchmarks

The majority of early benchmarks for LCMs are built based on real-world tasks that inherently encompass long contexts, such as long-document question-answering, document-level summarization, and conversation understanding (Li et al., 2023b; Shaham et al., 2023; An et al., 2023; GoodAI, 2024; Bai et al., 2023; Dong et al., 2023; Zhang et al., 2024a; Lee et al., 2024; Levy et al., 2024). However, given that real-world tasks manifest in assorted formats and utilize varied evaluation methodologies (Yang et al., 2024b; Shi et al., 2024), synthetic tasks are increasingly employed in long-context scenarios (Hsieh et al., 2024), allowing for custom definition into various types, thereby enabling controlled studies of model capabilities. For instance, retrieval-based tasks require LCMs to extract specific information from a long synthetic context (Kamradt, 2024; Mohtashami and Jaggi, 2023; Xiao et al., 2024a; Liu et al., 2024; Wang et al., 2024; Zhang et al., 2024b), many-shot in-context learning tasks require LCMs to comprehend and follow input examples (Agarwal et al., 2024; Bertsch et al., 2024), long-form reasoning tasks demand LCMs to respond based on clues within the long context (Kuratov et al., 2024; Karpinska et al., 2024). Nevertheless, recent works (Yen et al., 2024; Hsieh et al., 2024) have indicated that long-context benchmarks struggle to distinguish differences between LCMs with a limited testing set. At the same time, it remains unclear whether the models truly follow the contextual information when generating responses, which further leads to inconsistent LCM performance across different benchmarks. Therefore, we add an additional evaluation criterion, i.e., fidelity, to enable more effective and efficient assessments. Evaluating fidelity can better reflect whether LCMs respond based on the context, making the evaluation more universal and comprehensive.

2.2 Citation Generation

The citation generation task aims to evaluate the model's fidelity to the context by verifying whether its predictions are supported by the reference sources (Li et al., 2023a). Early works mainly focus on the evaluation perspective, aiming to more accurately assess the fidelity of models (Rashkin et al., 2023; Qian et al., 2023; Kamalloo et al., 2023; Li et al., 2023c) across different tasks and domains (e.g., single-document QA (Bohnet et al., 2022), fact checking (Honovich et al., 2022)) and domains (e.g., science (Funkquist et al., 2022), commerce (Liu et al., 2023)). With the advancement of generative AI, citation generation has begun to require models themselves to generate citations that support their predictions (Gao et al., 2023). More recently, Bai et al. (2024) introduced *LongCite*, which shares a similar idea with our work by extending citation generation to long-context question-answering tasks. Compared with

Tasks	Source	Evaluation Metric	Length Distribution						Total
			0~8k	8~16k	16~24k	24~32k	32~40k	40~48k	
Single-document QA (NarrativeQA*: 256, Natural Questions*: 256)									
NarrativeQA	(Kočiskỳ et al., 2018)	Prec., Rec.	40	40	40	40	40	40	240
Natural Questions	(Kwiatkowski et al., 2019)	Prec., Rec.	-	-	40	40	40	40	160
Multi-document QA (HotpotQA*: 128, 2WikiMultihopQA*: 128)									
HotpotQA	(Yang et al., 2018)	Prec., Rec.	40	40	40	40	40	40	240
2WikiMultihopQA	(Ho et al., 2020)	Prec., Rec.	40	40	40	40	40	40	240
Summarization (MultiNews*: 128, GovReport*: 128, QMSum*: 128)									
MultiNews	(Ghalandari et al., 2020)	Rouge-L	20	20	20	20	20	-	100
GovReport	(Huang et al., 2021)	Rouge-L	40	40	40	40	40	40	240
QMSum	(Zhong et al., 2021)	Rouge-L	20	20	20	20	-	-	80
Dialogue Understanding (LoCoMo*: 256, DialSim*: 256)									
LoCoMo	(Maharana et al., 2024)	Prec., Rec.	40	40	40	40	40	40	240
DialSim	(Kim et al., 2024)	Prec., Rec.	40	40	40	40	40	40	240
Synthetic Task (NIAH*: 256, Counting Stars*: 128)									
NIAH	(Kamradt, 2024)	Rouge-1	20	20	20	20	20	20	120
Counting Stars	(Song et al., 2024)	Accuracy	30	30	30	30	30	30	180

Table 1: Statistic of tasks in L-CiteEval benchmark. The citation chunk size for each task is denoted with *.

LongCite, L-CiteEval is (1) **more comprehensive**: it covers a wider range of tasks, supporting longer context lengths, and strictly categorizes tasks by length intervals; (2) **more reproducible**: the evaluation process relies on both automatic metrics and strong LLMs (e.g., GPT-4), ensuring more accurate and reproducible evaluation results; and (3) **more efficient**: the data distribution is well-designed in our benchmark, with a limited amount of testing data, it can reflect the model's overall performance

3 *L-CiteEval*: Task and Construction

3.1 Problem Definition

Given the long reference context T and question Q, the model is expected to generate the response R that contains both statements $\mathcal{S} = \{s_1, s_2, \cdots, s_n\}$ and their corresponding citations $\mathcal{C} = \{c_1, c_2, \cdots, c_n\}$. To facilitate citation generation by the model, we split the context T into chunks, assigning each chunk a unique citation index. The model then generates the corresponding index to indicate the chunks it references.

3.2 Benchmark Construction

There are 5 main categories in L-CiteEval benchmark: Single-Document QA, Multi-Document QA, Summarization, Dialogue Understanding, and Synthetic tasks. To ensure the accuracy of the evaluation data, we construct the benchmark mainly based on the existing short-context testing sets[2], which are commonly manually verified. We report the data source in Table 1. The construction process for each task consists of 3 steps, including (1) Seed Data & Padding Data Sampling, (2) Padding Data Filtering, and (3) Length Extension.

Step1: Seed Data & Padding Data Sampling Given the large volume of testing data in each open-source benchmark, we first select a subset $\mathcal{D}_{seed}$ from these benchmarks for subsequent processing and sample the padding data $\mathcal{D}_{pad}$ from the remaining testing datasets for length extension. We divide all the sampled data ($\mathcal{D}_{seed}$ and $\mathcal{D}_{pad}$) into chunks of approximately equal size, with sentences as the basic unit. Specifically, for tasks involving concentrated information, e.g., single-document QA, we employ smaller chunk sizes, while for tasks involving dispersed information, e.g., summarization, we use larger chunk sizes.

Step2: Padding Data Filtering Using $\mathcal{D}_{pad}$ to extend the length of a short-context dataset could potentially influence the model prediction. Therefore, we filter the padding data that might affect the predictions based on overlapping entities in the context. We apply spaCy (Honnibal and Montani, 2017), a NER model f_θ, to extract all the entities E from the reference context $\mathcal{T}_{seed}$ in $\mathcal{D}_{seed}$, as well as the entities from the reference context $\mathcal{T}'_{seed}$ in $\mathcal{D}_{pad}$. Then, we keep the padding data $\mathcal{D}^*_{pad}$ that share a small entity overlaps with those in $\mathcal{D}_{seed}$:

$$\begin{aligned}\mathcal{D}^*_{pad} = \{&\mathcal{D}'_{pad} \mid \mathcal{T}_{seed} \sim \mathcal{D}_{seed}, \mathcal{T}'_{seed} \sim \mathcal{D}_{pad}, \\ &|f_\theta(\mathcal{T}_{seed}) \cap f_\theta(\mathcal{T}'_{seed})| \leq \delta\},\end{aligned} \tag{1}$$

---

[2] Lengths of most samples in these datasets are within 12K.

Model	**#Param**	**Arch.**
GPT-4o (OpenAI, 2024a)	🔒	🔒
o1-mini (OpenAI, 2024b)	🔒	🔒
Claude-3.5-Sonnet (anthropic, 2024)	🔒	🔒
Qwen2.5-3B-Instruct (Team, 2024)	3B	Dec
Phi3.5-mini-instruct (Abdin et al., 2024)	3.8B	Dec
Llama3.1-8B-Instruct (Llama)	8B	Dec
GLM4-9B-Chat (GLM et al., 2024)	9B	Dec
Mistral-NeMo-Instruct (Mistral, 2024)	12B	Dec
Qwen2-57B-A14B-Instruct (Yang et al., 2024a)	57B	MoE
Llama3.1-70B-Instruct (Llama)	70B	Dec
Llama3-ChatQA-2-70B (Xu et al., 2024a)	70B	Dec

Table 2: Statistic of LCMs, where 🔒 denotes closed-sourced model and Dec denotes decoder-only model.

where δ is the threshold to control the entity overlap between $\mathcal{T}_{seed}$ and $\mathcal{T}'_{seed}$, and $\mathcal{D}_{seed} \cap \mathcal{D}^*_{pad} = \varnothing$. We set $\delta = 5$ to filter out padding data that may potentially impact the results.

Step3: Length Extension We leverage $\mathcal{D}^*_{pad}$ to extend the context length of $\mathcal{D}_{seed}$. Given the target length interval of each task, we first sort the data based on the original context length of each task and then randomly sample contexts from $\mathcal{D}^*_{pad}$ to fill in the missing target length intervals. To decouple the impact of task difficulty and content length on predictions, we introduce two variants: *L-CiteEval-Length* that assesses models from the context length perspective and *L-CiteEval-Hardness* that assesses models based on question difficulty. For L-CiteEval-Length, we use the same $\mathcal{D}_{seed}$ and different $\mathcal{D}^*_{pad}$ to extend to context length. For L-CiteEval-Hardness, we first quantify and rank the difficulty of each question based on the model prediction results[3]. Then, we categorize the difficulty into three levels: easy, medium, and hard, based on the response accuracy. We use the same $\mathcal{D}^*_{pad}$ to extend the context length for each difficulty level.

Benchmark Overview For clarity, we list the characteristics of three benchmarks below:

- ***L-CiteEval*** benchmark is designed to evaluate both fidelity and downstream task capabilities of LCMs regardless of question difficulty and context length. This benchmark comprises 2,080 test samples across 11 tasks of 5 categories, with context lengths ranging from 8K to 64K.
- ***L-CiteEval-Length*** benchmark is designed to evaluate models from the context length perspective, which is constructed with the same seed data (ensuring the same question difficulty) but different padding data (varying context length). This benchmark consists of 4 tasks across 4 categories, including NarrativeQA (Single-Doc QA), HotpotQA (Multi-Doc QA), GovReport (Summarization), and Counting Stars (Synthetic task), with each task containing 200 testing samples and 3 length intervals: 8K, 16K, and 32K.
- ***L-CiteEval-Hardness*** benchmark is designed to evaluate models from the task difficulty perspective, which is constructed with the different seed data (varying question difficulty) but the same padding data sources (same context). This benchmark shares the same data distribution and volume as L-CiteEval-Length, with the only difference being that the categorization is based on task difficulty (Easy, Medium, and Hard) rather than the context length.

3.3 Verification Pipeline

We assess LCMs from two aspects: generation quality and citation quality. For generation quality, we use evaluation metrics corresponding to specific downstream tasks, e.g., ROUGE for summarization tasks (Lin, 2004). For citation quality, following Gao et al. (2023), we adopt Citation Recall (`CR`) to measure whether the citations fully support the model's statements, Citation Precision (`CP`) to identify irrelevant citations, and F_1 score to reflect the overall citation performance. Additionally, we report Citation Number (`CN`) to indicate how many citations the model uses to support its statement. To automatically detect whether citations support the corresponding statements, we utilize the long-context NLI model DeBERTa-base-long-nli (Sileo, 2024) to better align with long-context scenarios. Apart from the aforementioned automatic evaluation metrics, we also leveraged strong LLMs for evaluation (Gu et al., 2024) to ensure the accuracy of the assessment. Details of citation metrics and evaluation process are shown in Appendix A.

4 Experiment

As shown in Table 2, we experiment with 11 latest cutting-edge LCMs, including 3 closed-source and 8 open-source models. Each model features a context window size of at least 128K tokens, with each possessing different model parameters (from 3B to 70B) and model architectures (decoder-only dense models and MoE models). We assess all the LCMs on L-CiteEval and then select 5 repre-

---

[3] We categorize the difficulty of each sample with GPT-4o since GPT-4o has been proven to exhibit the highest preference similarity with human annotators (Yadav et al., 2024).

Models	Single-Doc QA				Dialogue Understanding				Needle in a Haystack			
	CP	CR	F_1	N	CP	CR	F_1	N	CP	CR	F_1	N
🔒 ***Closed-source LCMs***												
GPT-4o	32.05	**38.12**	33.48	2.02	53.90	**64.25**	**56.76**	2.17	**82.08**	**82.50**	**82.22**	1.01
Claude-3.5-sonnet	**38.70**	37.79	**37.43**	3.54	**54.45**	50.48	51.45	2.83	73.33	76.67	74.31	1.10
o1-mini	29.83	35.33	31.66	3.38	45.54	50.74	47.21	2.63	28.47	30.83	29.17	1.46
🔓 ***Open-source LCMs***												
Qwen2.5-3b-Ins	7.13	5.83	6.00	1.75	9.53	9.71	8.41	2.33	12.08	12.50	12.22	1.04
Llama-3.1-8B-Ins	22.68	24.73	22.64	2.59	51.86	**57.58**	53.50	2.08	35.14	36.67	35.56	0.95
Glm-4-9B-chat	**29.00**	**28.66**	**28.05**	2.21	**54.54**	55.62	**53.58**	1.78	46.11	50.00	47.22	1.12
Qwen2-57B-A14B-Ins	4.90	3.43	3.82	1.27	22.63	22.54	21.61	1.80	15.83	15.83	15.83	1.10
Llama-3.1-70B-Ins	25.89	26.89	26.11	1.23	51.71	56.20	53.19	1.76	**54.17**	**54.17**	**54.17**	0.87

Table 3: Citation quality of LCMs in information-concentrated tasks in L-CiteEval.

Models	Multi-Doc QA				Summarization				Counting Stars			
	CP	CR	F_1	N	CP	CR	F_1	N	CP	CR	F_1	N
🔒 ***Closed-source LCMs***												
GPT-4o	57.48	**58.50**	56.10	1.71	34.37	54.28	41.60	22.86	**83.37**	**81.18**	**81.71**	4.54
Claude-3.5-sonnet	**66.85**	55.62	**58.58**	2.44	**36.70**	**55.03**	**43.45**	17.70	73.01	75.83	73.15	4.81
o1-mini	49.95	49.60	48.58	1.78	20.23	33.61	24.83	19.58	34.06	46.46	38.45	6.73
🔓 ***Open-source LCMs***												
Qwen2.5-3b-Ins	13.17	8.04	9.37	1.96	7.72	12.15	9.09	9.52	3.82	1.81	2.01	1.66
Llama3.1-8B-Ins	43.41	42.15	41.64	1.62	19.57	23.03	20.83	18.31	16.87	23.33	19.18	4.19
Glm4-9B-chat	47.91	44.75	45.09	1.64	**29.16**	**37.29**	**31.92**	11.38	18.15	16.04	16.21	4.52
Qwen2-57B-A14B-Ins	17.30	12.07	13.61	1.06	4.01	3.37	3.19	3.81	4.37	4.44	4.24	4.24
Llama3.1-70B-Ins	**49.64**	**54.02**	**50.74**	1.42	25.50	31.99	27.91	11.78	**66.85**	**61.74**	**63.73**	4.37

Table 4: Citation quality of LCMs in information-dispersed tasks in L-CiteEval.

sentative LCMs (including 1 closed-source LCMs and 4 open-source LCMs) to further evaluate on L-CiteEval-Length and L-CiteEval-Hardness datasets. We present the results of 3 open-source LCMs, with additional evaluation results, including LLM-based assessments and retrieval-based methods, in Appendix B, and provide the demonstration of prompt and the error analysis for each task in Appendix J.

4.1 Model Performance on L-CiteEval

We report citation quality in Tab. 3 (tasks that require models to extract information from several citation chunks) and Tab. 4 (tasks that require models to retrieve information from the entire context), and show the generation quality in Tab. 5. Notably, given the varying capability preferences of different models and the broad range of tasks covered by L-CiteEval, no single model can consistently achieve the best performance. For clarity, we use underlines to highlight our key insights.

4.1.1 Analysis of Citation Quality

Performance of Open-source LCMs There is significant room for open-source LCMs to improve and medium-sized LCMs (Llama3.1-8B-instruct and GLM4-9B-Chat) are highly competitive, with performance that matches or even exceeds that of LCMs with large parameters (Llama3.1-70B-instruct). Our key findings are: (1) citation quality does not consistently improve with an increase in model parameters. While large LCMs (70B) generally perform well, medium-sized models (8B and 9B) deliver surprisingly strong results; (2) the effective activated parameters are critical. For instance, the MoE LCM (Qwen2-57B-A14B) demonstrates poorer citation quality, even underperforming smaller dense models like Llama3.1-8B.

Performance of Closed-source LCMs Among closed-source LCMs, GPT-4o and Claude-3.5-sonnet show exceptional performance, with GPT-4o surpassing all the open-source LCMs in citation quality across all tasks. Notably, while o1-mini achieves unmatched results in reasoning tasks such as GSM8K (Cobbe et al., 2021) and Livecodebench (Jain et al., 2024), its citation generation performance declines significantly in long-context scenarios. Specifically, in synthetic and

Models	Single-Doc QA		Multi-Doc QA		Summ.	Dialogue		Synthetic	
	Prec.	Rec.	Prec.	Rec.	Rouge-L	Prec.	Rec.	Rouge-1†	Acc‡
🔒 ***Closed-source LCMs***									
GPT-4o	**11.78**	70.37	**10.34**	**87.38**	20.15	**9.81**	**65.35**	**96.25**	**91.88**
Claude-3.5-sonnet	5.96	**71.96**	4.30	80.77	**22.06**	3.71	57.80	94.46	69.65
o1-mini	10.30	66.44	7.36	64.25	19.22	7.02	54.27	56.52	57.29
🔓 ***Open-source LCMs***									
Qwen2.5-3b-Ins	8.91	60.28	3.82	52.41	22.39	4.58	40.77	84.06	26.81
Llama-3.1-8B-Ins	10.11	**68.13**	7.66	68.84	20.90	11.07	58.84	85.34	33.75
Glm-4-9B-chat	11.22	67.25	7.88	**77.97**	21.42	7.69	51.25	87.99	58.82
Qwen2-57B-A14B-Ins	12.93	61.71	15.25	57.53	**22.95**	14.32	52.23	94.20	63.61
Llama-3.1-70B-Ins	15.23	67.08	12.50	76.40	22.29	19.62	**62.91**	**94.58**	**89.03**

Table 5: Generation quality of LCMs on L-CiteEval, where † denotes the NIAH results, ‡ denotes the Counting Stars results, and Summ. denotes the summarization task.

summarization tasks that require LCMs to extract dispersed key information and effectively utilize retrieval data for response, o1-mini's performance falls markedly behind strong open-source models like Llama3.1-70B-instruct.

Open-source LCMs vs. Closed-source LCMs Overall, there is still a significant performance gap between open-source LCMs and closed-source LCMs (excluding o1-mini), especially in tasks involving reasoning. Specifically, we can observe that: (1) closed-source LCMs generally provide more accurate citations with larger F_1 score and tend to leverage more citation chunks (larger `N`) to support the statement; (2) for tasks involving reasoning, such as *Counting Stars* synthetic task that requires LCM to retrieve and count specific tokens from the long context, although strong open-source LCMs like GLM4-9B-Instruct cite a comparable number of segments to their closed-source counterparts, the citation quality is notably lower, leading to a performance gap of nearly 20 F_1 points.

4.1.2 Analysis of generation quality

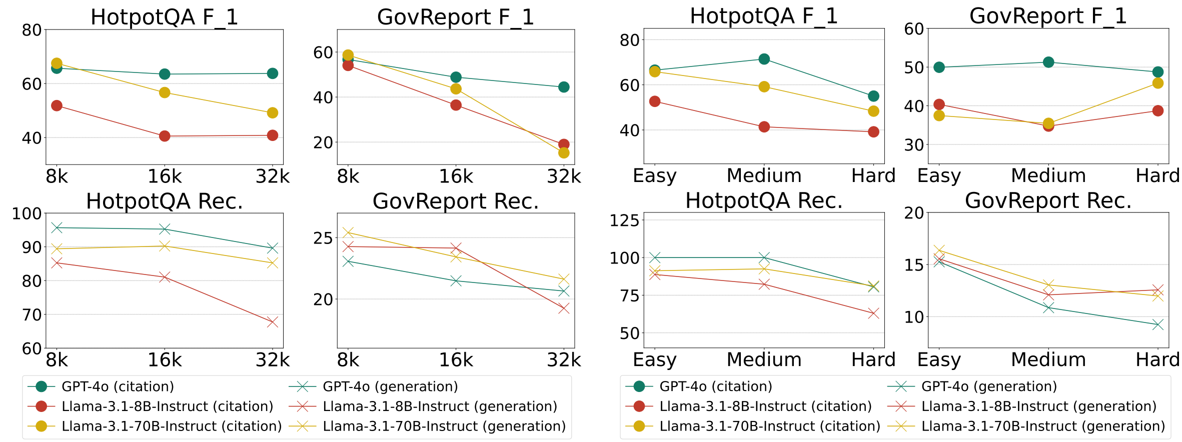
From Table 5, we observe that in Single-Doc QA, Multi-Doc QA, and Dialogue Understanding tasks, closed-source LCMs significantly outperform open-source LCMs in terms of recall scores. However, closed-source models exhibit notably low accuracy. Based on our error analysis in Appendix J, we find that closed-source models tend to produce overly verbose statements to justify their results, which ultimately leads to lower precision scores. In Summarization and Synthetic tasks, the performance gap between closed-source and strong open-source LCMs narrows, as evidenced by close evaluation results, such as the 22.06 Rouge-L score of Claude-3.5-sonnet compared to the 22.95 Rouge-L score of Qwen2-57B-A14B-Instruct in summarization tasks. Besides, open-source LCMs tend to demonstrate better performance as the number of model parameters increases. However, combined with the aforementioned lackluster citation quality of large LCMs, we hypothesize that large LCMs rely heavily on their internal knowledge (which may include task-specific knowledge) rather than responding based on the provided context. This finding is also consistent with (Intel, 2024).

4.2 Controlled Study on L-CiteEval

We evaluate LCMs on L-CiteEval-Length and L-CiteEval-Hardness. More experiment details and evaluation results are shown in Appendix D.

4.2.1 Impact of Context Length

We present the model performance on L-CiteEval-Length in Fig. 2(a). When keeping task difficulty constant but progressively extending the context length, we observe a decline in open-source model performance. Notably, the smallest model, Llama3.1-8B-Instruct, is the most adversely affected by longer contexts. For instance, in *HotpotQA* task, its F_1 score drops by approximately 10 points as the context length increases from 8K to 32K. Larger models like Llama3.1-70B-Instruct, demonstrate greater robustness, with only minor performance degradation. In contrast, closed-source LCM (GPT-4o) displays remarkable stability, showing minimal performance decline even with extended contexts. These findings indicate that open-source LCMs are more vulnerable to irrelevant contextual information, leading to a notable decline in both fidelity and generation quality.



(a) Model Performance on L-CiteEval-Length. (b) Model Performance on L-CiteEval-Hardness.

Figure 2: Model Performance on L-CiteEval-Length benchmark and L-CiteEval-Hardness benchmark, where we apply F_1 metric to assess citation quality and recall score (`Rec.`) to assess generation quality.

4.2.2 Impact of Task Difficulty

We show the model performance on L-CiteEval-Hardness benchmark in Fig. 2(b), where we can observe that as task difficulty increases, the generation quality (`Rec.` score) of LCMs generally declines. However, citation quality does not follow a clear trend, which underscores a gap between citation quality and downstream task performance. This aligns with our intuition that fidelity is not correlated with task difficulty, as the model can leverage its internal knowledge to answer questions of varying difficulty, rather than solely relying on the provided context.

5 Ablation Study

In this section, we investigate the effect of commonly used methods in the long-context field, including context compression methods (Verma, 2024) and the inference efficiency methods (Xiao et al., 2024c), on model fidelity in § 5.1. Then, we analyze the benefits brought by citation generation in § 5.2 and reveal the relationship between the explicit model citation process and model's implicit information retrieval mechanism in § 5.3.

5.1 Effectiveness of Context Compression and Inference Efficiency Methods

There are two mainstream context compression methods in the long-context scenario: context compression via summarization (Xu et al., 2024b; Jha et al., 2024) and retrieval-based methods (RAG) (Leng et al., 2024; Li et al., 2024c; Yu et al., 2024). For the **summarization-based** method, to ensure the integrity of citation chunks, we employ the Llama3.1-70B-Instruct model to summarize each chunk individually and concatenate the summarized chunks as the model's new input. For the **retrieval-based** method, we leverage the dense retriever GTR-T5-XXL (Ni et al., 2021) to identify citation chunks relevant to the question and select the top 32 citation segments with the highest retrieval scores as the model's new input. We also test with two inference efficiency methods: StreamingLLM (Xiao et al., 2024c) and DuoAttention (Xiao et al., 2024b).

Context Compression Result As shown in Tab. 6, we present the performance of two Llama3.1 models with different parameters (8B and 70B) and compare them with GPT-4o. We observe that for the Single-Doc QA task (i.e., Natural Questions), context compression methods can significantly enhance the citation quality of LCMs, with the Llama3.1-70B-Instruct model greatly outperforming GPT-4o in the Natural Question task (51.60 vs. 36.44 of `F_1` score). However, for the Multi-Doc QA task (i.e., HotpotQA), these methods compromise model's fidelity. For generation quality, context compression methods show side effects, where details may be omitted due to context compression. More experimental results are shown in Appendix E.1 and E.2.

Inference Efficiency Result As shown in Tab. 7, we find that Llama-2-7b-chat is too weak to handle L-CiteEval task[4]. For Llama-3-8B-Instruct model, although DuoAttention can significantly improve the model's inference efficiency, it significantly

---

[4] Since StreamingLLM code only supports LLama2.

Model	Natural Questions			HotpotQA		
	`F_1`	`Rec.`	`Ctx.`	`F_1`	`Rec.`	`Ctx.`
GPT-4o	36.44	**82.41**	-	61.81	**90.63**	-
LLama3.1-8B-Ins	21.96	81.93	35,039	40.77	78.30	28,080
+ Summarization	46.36	59.90	7,078	39.23	53.23	11,654
+ Retrieval	30.29	76.74	9,983	49.80	78.93	5,327
Llama3.1-70B-Ins	25.13	76.54	35,039	54.86	85.39	28,080
+ Summarization	**60.97**	68.14	7,078	47.50	59.31	11,654
+ Retrieval	51.60	80.98	9,983	**62.22**	81.08	5,327

Table 6: Model Performance with context compression methods, where we report `F_1` for citation performance, `Rec.` for generation quality, and the average context length `Ctx.` for each method.

Models	NarrativeQA		GovReport		LoCoMo	
	`F_1`	`Rec.`	`F_1`	`Rouge-L`	`F_1`	`Rec.`
Llama-2-7b-chat	0.00	0.00	0.00	0.00	0.00	0.00
+ StreamingLLM	0.00	0.00	0.00	0.00	0.00	0.00
Llama-3-8B-Ins	20.17	63.76	0.83	25.44	11.88	61.51
+ DuoAttention	7.95	61.96	0.00	25.43	7.41	47.98

Table 7: Citation quality and generation quality of long-context inference efficiency methods.

degrades the performance. More implementation details and results of long-context inference efficiency are shown in Appendix E.3.

5.2 Benefit of Citation Generation Process

As shown in Tab. 8, we can find that model response with citation can boost both the model performance and its fidelity. This can be attributed to the LCM performing additional reasoning steps, i.e., leveraging evidence within the context to support its statements, which has been proven to benefit the model's peformance (Li et al., 2024a,b). More results are shown in Appendix F.

5.3 Analysis of Model Implicit Information Retrieval Mechanism

We then investigate why generating citations can improve generation quality by analyzing the model's implicit information retrieval mechanism (Wu et al., 2024). Specifically, we calculate the attention scores on the critical chunks to reflect whether the model focuses on those pieces of evidence. We conduct the experiments on HotpotQA with two strong LCMs, including Llama-3.1-8B-Instruct and GLM-4-9B-Chat. As shown in Fig. 3, each dot in the figure represents the number of citations generated by the model and the number of citations attended to by the model's attention mechanism. Ideally, if the model can accurately output all citations attended to by its attention mechanism, all the dots would align along the diagonal green curve. We plot the correlation coefficient (r) between the number of generated citations and those retrieved by the attention mechanism, finding all the correlation values exceed 0.7. However, when the model does not include citations in its output, the corresponding correlation coefficients indicate that the model struggles to detect citations effectively. More implementation details and results can be found in Appendix G.

Model	Single-Doc QA		Multi-Doc QA		Summ.
	`Prec.`	`Rec.`	`Prec.`	`Rec.`	`R-L`
GPT-4o	11.78	70.37	10.34	**87.38**	**20.15**
w/o citation	**12.18**	**70.59**	**11.09**	85.09	19.00
LLama3.1-8B-Ins	10.11	**68.13**	**7.66**	**68.84**	**20.90**
w/o citation	**10.56**	64.83	4.46	55.00	18.40
Glm-4-9B-chat	**11.22**	**67.25**	**7.88**	**77.97**	**21.42**
w/o citation	8.27	66.85	6.55	71.25	18.35

Table 8: Generation quality of LCMs with citations (default, gray background) and without citations.

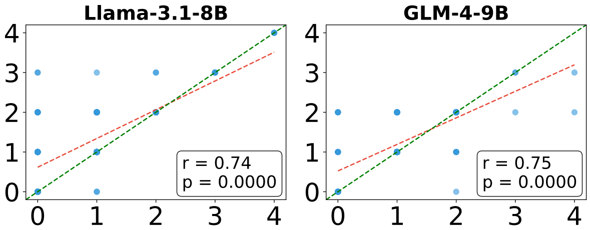


Figure 3: Pearson correlation analysis between generated citations and implicit retrieval mechanisms: the x-axis shows the number of correct generated citations, and the y-axis shows the number of citations attended by the attention. The red curve represents the fitted correlation and the green curve indicates the best correlation.

6 Conclusion

In this paper, we introduce L-CiteEval, an out-of-the-box evaluation suite featuring a multi-task long-context benchmark and a corresponding evaluation pipeline. The benchmark includes 5 major task categories spanning 11 long-context tasks, with context lengths ranging from 8K to 48K. Comprehensive testing across 11 state-of-the-art LCMs reveals that open-source LCMs often rely on intrinsic knowledge rather than the provided context to generate responses. Moreover, we find that context compression and inference efficiency methods albeit with the trade-offs between generation (efficiency or performance) and citation. Finally, we uncover a correlation between citation generation and the implicit information retrieval mechanism of LCMs, highlighting the benefits of citation generation in long-context tasks.

Limitation

In this paper, we introduce L-CiteEval. Compared to existing long-context benchmarks, L-CiteEval includes an additional evaluation dimension, i.e., fidelity, which is a crucial property for LCMs. With limited tasks and a range of context lengths, we can significantly reflect the capability of the model. However, there are still some limitations:

- Currently, many benchmarks are facing serious data leakage issues (Apicella et al., 2024), which is not just a problem in the long-text evaluation domain but across the entire evaluation field. An effective solution is to continuously update the testing data through anonymous submissions to prevent data leakage. Therefore, in our future work, we will continue to refine L-CiteEval by creating an anonymous system where we dynamically adjust tasks and data to mitigate the risk of data leakage.

- Currently, the data in L-CiteEval is still limited. While we believe that using less data can enhance evaluation efficiency, it can also lead to potential issues with data distribution bias. There is a trade-off between the comprehensiveness of the evaluation and efficiency, and in this paper, L-CiteEval emphasizes efficiency. Therefore, in future work, we will propose another version, an ***L-CiteEval-Ultra*** version, which will cover a broader range of data distributions and longer context lengths to provide a more comprehensive evaluation of LCMs.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Rishabh Agarwal, Avi Singh, Lei M Zhang, Bernd Bohnet, Stephanie Chan, Ankesh Anand, Zaheer Abbas, Azade Nova, John D Co-Reyes, Eric Chu, et al. 2024. Many-shot in-context learning. *arXiv preprint arXiv:2404.11018*.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2023. L-eval: Instituting standardized evaluation for long context language models. *arXiv preprint arXiv:2307.11088*.

anthropic. 2024. Claude-3-5-sonnet model card. *blog*.

Andrea Apicella, Francesco Isgrò, and Roberto Prevete. 2024. Don't push the button! exploring data leakage risks in machine learning and transfer learning. *arXiv preprint arXiv:2401.13796*.

Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, Juanzi Li, et al. 2024. Longcite: Enabling llms to generate fine-grained citations in long-context qa. *arXiv preprint arXiv:2409.02897*.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

Amanda Bertsch, Maor Ivgi, Uri Alon, Jonathan Berant, Matthew R Gormley, and Graham Neubig. 2024. In-context learning with long-context models: An in-depth exploration. *arXiv preprint arXiv:2405.00200*.

Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, et al. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Bamboo: A comprehensive benchmark for evaluating long text modeling capacities of large language models. *arXiv preprint arXiv:2309.13345*.

Martin Funkquist, Ilia Kuznetsov, Yufang Hou, and Iryna Gurevych. 2022. Citebench: A benchmark for scientific citation text generation. *arXiv preprint arXiv:2212.09577*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627*.

Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. 2020. A large-scale multi-document summarization dataset from the wikipedia current events portal. *arXiv preprint arXiv:2005.10070*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

GoodAI. 2024. Introducing goodai ltm benchmark. *blog*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. True: Re-evaluating factual consistency evaluation. *arXiv preprint arXiv:2204.04991*.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. *arXiv preprint arXiv:2104.02112*.

Intel. 2024. Do smaller models hallucinate more? *blog*.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*.

Siddharth Jha, Lutfi Eren Erdogan, Sehoon Kim, Kurt Keutzer, and Amir Gholami. 2024. Characterizing prompt compression methods for long context inference. *arXiv preprint arXiv:2407.08892*.

Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. Hagrid: A human-llm collaborative dataset for generative information-seeking with attribution. *arXiv preprint arXiv:2307.16883*.

Gregory Kamradt. 2024. Needle in a haystack - pressure testing llms. *Github*.

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. One thousand and one pairs: A" novel" challenge for long-context language models. *arXiv preprint arXiv:2406.16264*.

Jiho Kim, Woosog Chay, Hyeonji Hwang, Daeun Kyung, Hyunseung Chung, Eunbyeol Cho, Yohan Jo, and Edward Choi. 2024. Dialsim: A real-time simulator for evaluating long-term dialogue understanding of conversational agents. *arXiv preprint arXiv:2406.13144*.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. 2024. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. *arXiv preprint arXiv:2406.10149*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Jinhyuk Lee, Anthony Chen, Zhuyun Dai, Dheeru Dua, Devendra Singh Sachan, Michael Boratko, Yi Luan, Sébastien MR Arnold, Vincent Perot, Siddharth Dalmia, et al. 2024. Can long-context language models subsume retrieval, rag, sql, and more? *arXiv preprint arXiv:2406.13121*.

Quinn Leng, Jacob Portes, Sam Havens, Matei Zaharia, and Michael Carbin. 2024. Long context rag performance of large language models. *arXiv preprint arXiv:2411.03538*.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. *arXiv preprint arXiv:2402.14848*.

Dongfang Li, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Ziyang Chen, Baotian Hu, Aiguo Wu, and Min Zhang. 2023a. A survey of large language models attribution. *arXiv preprint arXiv:2311.03731*.

Huayang Li, Pat Verga, Priyanka Sen, Bowen Yang, Vijay Viswanathan, Patrick Lewis, Taro Watanabe, and Yixuan Su. 2024a. Alr2: A retrieve-then-reason framework for long-context question answering. *arXiv preprint arXiv:2410.03227*.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2023b. Loogle: Can long-context language models understand long contexts? *arXiv preprint arXiv:2311.04939*.

Xinze Li, Yixin Cao, Liangming Pan, Yubo Ma, and Aixin Sun. 2023c. Towards verifiable generation: A benchmark for knowledge-aware language model attribution. *arXiv preprint arXiv:2310.05634*.

Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. 2024b. Chain of thought empowers transformers to solve inherently serial problems. *arXiv preprint arXiv:2402.12875*.

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024c. Retrieval augmented generation or long-context llms? a comprehensive study and hybrid approach. *arXiv preprint arXiv:2407.16833*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848*.

Meta Introducing Llama. 3.1: Our most capable models to date.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of llm agents. *arXiv preprint arXiv:2402.17753*.

Supriya Manna and Niladri Sett. 2024. Faithfulness and the notion of adversarial sensitivity in nlp explanations. *arXiv preprint arXiv:2409.17774*.

Mistral. 2024. Mistral nemo. *blog*.

Amirkeivan Mohtashami and Martin Jaggi. 2023. Landmark attention: Random-access infinite context length for transformers. *arXiv preprint arXiv:2305.16300*.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. *arXiv preprint arXiv:2305.16938*.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y Zhao, Yi Luan, Keith B Hall, Ming-Wei Chang, et al. 2021. Large dual encoders are generalizable retrievers. *arXiv preprint arXiv:2112.07899*.

Shiwen Ni, Xiangtao Kong, Chengming Li, Xiping Hu, Ruifeng Xu, Jia Zhu, and Min Yang. 2024. Training on the benchmark is not all you need. *arXiv preprint arXiv:2409.01790*.

OpenAI. 2024a. Gpt-4o model card. *blog*.

OpenAI. 2024b. o1-mini model card. *blog*.

Hongjing Qian, Yutao Zhu, Zhicheng Dou, Haoqi Gu, Xinyu Zhang, Zheng Liu, Ruofei Lai, Zhao Cao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Webbrain: Learning to generate factually correct articles for queries by grounding on large web corpus. *arXiv preprint arXiv:2304.04358*.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Jiayang Cheng, Cunxiang Wang, Shichao Sun, Huanyu Li, et al. 2024. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. *arXiv preprint arXiv:2408.08067*.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. Zeroscrolls: A zero-shot benchmark for long text understanding. *arXiv preprint arXiv:2305.14196*.

Dan Shi, Renren Jin, Tianhao Shen, Weilong Dong, Xinwei Wu, and Deyi Xiong. 2024. Ircan: Mitigating knowledge conflicts in llm generation via identifying and reweighting context-aware neurons. *arXiv preprint arXiv:2406.18406*.

Damien Sileo. 2024. tasksource: A large collection of NLP tasks with a structured dataset preprocessing framework. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15655–15684, Torino, Italia. ELRA and ICCL.

Mingyang Song, Mao Zheng, and Xuan Luo. 2024. Counting-stars: A multi-evidence, position-aware, and scalable benchmark for evaluating long-context large language models. *Preprint*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Sourav Verma. 2024. Contextual compression in retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2409.13385*.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2024. Augmenting language models with long-term memory. *Advances in Neural Information Processing Systems*, 36.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2024. Retrieval head mechanistically explains long-context factuality. *arXiv preprint arXiv:2404.15574*.

Chaojun Xiao, Pengle Zhang, Xu Han, Guangxuan Xiao, Yankai Lin, Zhengyan Zhang, Zhiyuan Liu, Song Han, and Maosong Sun. 2024a. Infllm: Unveiling the intrinsic capacity of llms for understanding extremely long sequences with training-free memory. *arXiv preprint arXiv:2402.04617*.

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. 2024b. Duoattention: Efficient long-context llm inference with retrieval and streaming heads. *arXiv preprint arXiv:2410.10819*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024c. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*.

Peng Xu, Wei Ping, Xianchao Wu, Zihan Liu, Mohammad Shoeybi, and Bryan Catanzaro. 2024a. Chatqa 2: Bridging the gap to proprietary llms in long context and rag capabilities. *arXiv preprint arXiv:2407.14482*.

Yang Xu, Yunlong Feng, Honglin Mu, Yutai Hou, Yitong Li, Xinghao Wang, Wanjun Zhong, Zhongyang Li, Dandan Tu, Qingfu Zhu, et al. 2024b. Concise and precise context compression for tool-using language models. *arXiv preprint arXiv:2407.02043*.

Sachin Yadav, Tejaswi Choppa, and Dominik Schlechtweg. 2024. Towards automating text annotation: A case study on semantic proximity annotation using gpt-4. *arXiv preprint arXiv:2407.04130*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Sohee Yang, Nora Kassner, Elena Gribovskaya, Sebastian Riedel, and Mor Geva. 2024b. Do large language models perform latent multi-hop reasoning without exploiting shortcuts? *arXiv preprint arXiv:2411.16679*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. 2024. Helmet: How to evaluate long-context language models effectively and thoroughly. *arXiv preprint arXiv:2410.02694*.

Tan Yu, Anbang Xu, and Rama Akkiraju. 2024. In defense of rag in the era of long-context language models. *arXiv preprint arXiv:2409.01666*.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, et al. 2024a. Infinitebench: Extending long context evaluation beyond 100k tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15262–15277.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, et al. 2024b. Infty bench: Extending long context evaluation beyond 100k tokens. *arXiv preprint arXiv:2402.13718*.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, et al. 2021. Qmsum: A new benchmark for query-based multi-domain meeting summarization. *arXiv preprint arXiv:2104.05938*.

A Details of Verification Pipeline

Before we calculate CR and CP metrics, we employ two models to identify the golden cited chunks within the context: the open-source NLI model deberta-base-long-nli[5] that provides a lightweight approach and GPT-4o[6] that provides a strong information extraction capability.

A.1 Calculation of CR and CP

Citation Recall CR CR measures whether all cited chunks fully support a given statement. For a statement s_i and its supported evidence $\mathcal{E}_i = \{e_{i,j}\}_{j=1}^{N}$, the evidence $e_{i,j}$ are concatenated into a whole passage P_i. Then $f_\theta(\cdot)$ is adopted to verify if P_i entails s_i, which can be implemented by a NLI model or GPT-4o. The calculation process of CR can be written as:

$$\texttt{CR} = \frac{\sum_{i=1}^{M} \mathbb{I}\left(f_\theta(P_i, s_i)\right)}{M},$$

where $\mathbb{I}(\cdot)$ denote whether P entails s_i and M denote the number of statements in a data instance.

Citation Precision CP CP evaluates the relevance of individual cited chunks by identifying "irrelevant" citations. For each evidence $e_{i,j} \in \mathcal{E}_i$, we remove $e_{i,j}$ from $\mathcal{E}_i$, forming a new set $\mathcal{E}'_{i,j}$. The evidences in $\mathcal{E}'_{i,j}$ are concatenated into $P'_{i,j}$, and the evaluation model is used to verify if $P'_{i,j}$ still supports the statement s_i. If removing $e_{i,j}$ does not affect the entailment, $e_{i,j}$ is considered irrelevant. CP is calculated as:

$$\texttt{CP} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} \mathbb{I}\left(f_\theta(P'_{i,j}, s_i)\right)}{N * M}.$$

B Full Evaluation Results on L-CiteEval

In this section, we present the results of 3 models that were not reported in the main text (i.e., Phi3.5-mini-Instruct, Mistral-Nemo-Instruct, and ChatQA2-70B), along with additional evaluation metrics, including LLM-based and retrieval-based evaluation results. As shown in Tab. 9, Tab. 10 and Tab. 11, we present the full evaluation of 11 LCMs, where we use the same metrics as those in the main text. Then, following appendix A, we calculate the citation quality with GPT-4o, and report the results in Tab. 12, and the corresponding instruction for evaluation is provided in Fig. 4.

The evaluation results from both the NLI model and GPT-4o exhibit a consistent ranking trend across various models and tasks. For instance, in the Single-Document QA task, closed-source models like Claude-3.5-sonnet and GPT-4o consistently outperform open-source models such as Qwen2.5-3b-Instruct and Qwen2-57B-A14B-Instruct in both CP and CR metrics across both evaluation methods. Similarly, in tasks like Dialogue Understanding and Multi-Document QA, closed-source models generally achieve higher citation quality scores compared to open-source models, regardless of whether the evaluation was conducted using the NLI model or GPT-4o. This alignment in model performance rankings suggests that both evaluation methods reliably differentiate between strong and weak models.

Despite the consistent ranking trends, there are noticeable differences in the absolute scores reported by the NLI model and GPT-4o. Typically, the GPT-4o evaluations yield lower CP, CR, and F_1 scores compared to the NLI model across most tasks and models. For example, GPT-4o rates Claude-3.5-sonnet with an F_1 score of 30.10 in Single-Document QA, whereas the NLI model assigns it a higher F_1 score of 37.43. This discrepancy indicates that while both evaluators agree on the relative performance of the models, they differ in their sensitivity or strictness in assessing citation quality. The NLI model may be more lenient, possibly due to differences in interpretative criteria or the inherent capabilities of the evaluation models.

The consistent rankings of the model performance indicate that the more cost-effective NLI model can reliably identify top-performing models, making it particularly suitable for large-scale evaluations where resources are constrained.

C Results with RAGChecker

In addition, we utilize RAGChecker (Ru et al., 2024) to evaluate the generation quality of the model. Specifically, we evaluate the model's response from two aspects: faithfulness and the recall score of correct claims in model response. As shown in Tab. 13, close-source models maintain higher faithfulness scores, indicating that their responses are more reliably grounded in the relevant context chunks. For example, GPT-4o and Claude-3.5-sonnet exhibit high recall and faithfulness across multiple tasks, ensuring that their answers are not only correct but also well-supported by the provided context. Conversely, though open-

---

[5] https://huggingface.co/tasksource/deberta-base-long-nli

[6] gpt-4o-2024-05-13

Models	Single-Doc QA				Dialogue Understanding				Needle in a Haystack			
	CP	CR	F_1	N	CP	CR	F_1	N	CP	CR	F_1	N
🔒 ***Closed-source LCMs***												
GPT-4o	32.05	**38.12**	33.48	2.02	53.90	**64.25**	**56.76**	2.17	**82.08**	**82.50**	**82.22**	1.01
Claude-3.5-sonnet	**38.70**	37.79	**37.43**	3.54	**54.45**	50.48	51.45	2.83	73.33	76.67	74.31	1.10
o1-mini	29.83	35.33	31.66	3.38	45.54	50.74	47.21	2.63	28.47	30.83	29.17	1.46
🔓 ***Open-source LCMs***												
Qwen2.5-3b-Ins	7.13	5.83	6.00	1.75	9.53	9.71	8.41	2.33	12.08	12.50	12.22	1.04
Phi-3.5-mini-Ins	21.06	20.46	19.14	2.86	20.39	24.27	20.57	2.27	11.67	12.50	11.94	1.08
Llama-3.1-8B-Ins	22.68	24.73	22.64	2.59	<u>51.86</u>	**57.58**	<u>53.50</u>	2.08	35.14	36.67	35.56	0.95
Glm-4-9B-chat	**29.00**	**28.66**	**28.05**	2.21	**54.54**	55.62	**53.58**	1.78	<u>46.11</u>	<u>50.00</u>	<u>47.22</u>	1.12
Mistral-Nemo-Ins	4.34	3.68	3.76	0.68	23.91	24.33	23.50	1.35	10.69	11.67	10.97	1.08
Qwen2-57B-A14B-Ins	4.90	3.43	3.82	1.27	22.63	22.54	21.61	1.80	15.83	15.83	15.83	1.10
Llama-3.1-70B-Ins	<u>25.89</u>	<u>26.89</u>	<u>26.11</u>	1.23	51.71	<u>56.20</u>	53.19	1.76	**54.17**	**54.17**	**54.17**	0.87
ChatQA-2-70B	21.75	22.54	21.92	1.12	47.67	51.25	48.77	1.29	39.17	39.17	39.17	0.75

Table 9: Citation quality of LCMs in information-concentrated tasks within L-CiteEval.

Models	Multi-Doc QA				Summarization				Counting Stars			
	CP	CR	F_1	N	CP	CR	F_1	N	CP	CR	F_1	N
🔒 ***Closed-source LCMs***												
GPT-4o	57.48	**58.50**	56.10	1.71	34.37	54.28	41.60	22.86	**83.37**	**81.18**	**81.71**	4.54
Claude-3.5-sonnet	**66.85**	55.62	**58.58**	2.44	**36.70**	**55.03**	**43.45**	17.70	73.01	75.83	73.15	4.81
o1-mini	49.95	49.60	48.58	1.78	20.23	33.61	24.83	19.58	34.06	46.46	38.45	6.73
🔓 ***Open-source LCMs***												
Qwen2.5-3b-Ins	13.17	8.04	9.37	1.96	7.72	12.15	9.09	9.52	3.82	1.81	2.01	1.66
Phi-3.5-mini-Ins	11.89	10.25	10.53	1.71	10.90	10.94	9.60	8.23	4.19	4.31	4.09	3.48
Llama-3.1-8B-Ins	43.41	42.15	41.64	1.62	19.57	23.03	20.83	18.31	16.87	<u>23.33</u>	<u>19.18</u>	4.19
Glm-4-9B-chat	<u>47.91</u>	44.75	45.09	1.64	**29.16**	**37.29**	**31.92**	11.38	<u>18.15</u>	16.04	16.21	4.52
Mistral-Nemo-Ins	17.61	15.45	15.85	0.70	11.21	14.85	12.40	5.45	3.09	3.68	3.26	2.32
Qwen2-57B-A14B-Ins	17.30	12.07	13.61	1.06	4.01	3.37	3.19	3.81	4.37	4.44	4.24	4.24
Llama-3.1-70B-Ins	**49.64**	**54.02**	**50.74**	1.42	<u>25.50</u>	<u>31.99</u>	<u>27.91</u>	11.78	**66.85**	**61.74**	**63.73**	4.37
ChatQA-2-70B	47.20	<u>49.51</u>	<u>47.92</u>	1.10	19.57	23.60	20.89	11.81	14.02	12.78	13.22	3.49

Table 10: Citation quality of LCMs in information-dispersed tasks within L-CiteEval.

source models like Glm-4-9B-chat and Llama-3.1-70B-Instruct demonstrate competitive but slightly lower faithfulness compared to closed-source models, most of the open-source models show lower faithfulness and recall, suggesting that their generated claims are less consistently supported by the relevant context. This disparity highlights the essential role of robust citation practices in achieving faithful and correct responses, further validating the interconnectedness of faithfulness and answer correctness in LCM performance. We also notice that open-source models like ChatQA-2-70B exhibit notable correctness in Tab. 11 but lower faithfulness in Tab. 13. These correct but unverifiable answers pose the challenge that the inability of the model to accurately attribute claims to specific chunks of the context undermines trustworthiness. Even if the answer is correct, the lack of a clear citation chain makes it impossible for users to verify the response, reducing its utility in critical applications. Worse still, if the model generates a hallucinated answer, it becomes harder to discern errors, as the incorrect information is presented with the same fluency as a correct response.

D Controlled Study of LCMs

We assess 5 representative LCMs with L-CiteEval-Length and L-CiteEval-Hardness and report the evaluation results in Tab. 15. We utilize `F_1` to reflect the citation quality and corresponding evaluation metrics to reflect the generation quality (Recall score for NarrativeQA, HotpotQA and LoCoMo tasks, Rouge-L for GovReport task, and Accuracy for Counting stars task).

Models	Single-Doc QA		Multi-Doc QA		Summ.	Dialogue		Synthetic	
	Prec.	Rec.	Prec.	Rec.	Rouge-L	Prec.	Rec.	Rouge-1†	Acc‡
🔒 ***Closed-source LCMs***									
GPT-4o	**11.78**	70.37	**10.34**	**87.38**	20.15	**9.81**	**65.35**	**96.25**	**91.88**
Claude-3.5-sonnet	5.96	**71.96**	4.30	80.77	**22.06**	3.71	57.80	94.46	69.65
o1-mini	10.30	66.44	7.36	64.25	19.22	7.02	54.27	56.52	57.29
🔓 ***Open-source LCMs***									
Qwen2.5-3b-Ins	8.91	60.28	3.82	52.41	22.39	4.58	40.77	84.06	26.81
Phi-3.5-mini-Ins	8.62	62.34	7.82	64.54	19.48	11.39	52.77	79.52	61.32
Llama-3.1-8B-Ins	10.11	**68.13**	7.66	68.84	20.90	11.07	58.84	85.34	33.75
Glm-4-9B-chat	11.22	67.25	7.88	**77.97**	21.42	7.69	51.25	87.99	58.82
Mistral-Nemo-Ins	10.53	59.71	8.78	67.70	20.83	9.27	49.26	90.01	18.06
Qwen2-57B-A14B-Ins	12.93	61.71	15.25	57.53	**22.95**	14.32	52.23	94.20	63.61
Llama-3.1-70B-Ins	15.23	67.08	12.50	76.40	22.29	19.62	**62.91**	**94.58**	**89.03**
ChatQA-2-70B	**43.25**	61.20	**34.95**	55.64	22.06	**26.57**	58.34	79.00	78.68

Table 11: Generation quality of LCMs on L-CiteEval, where † denotes the NIAH results, ‡ denotes the Counting Stars results, and Summ. denotes the summarization task.

Models	Single-Doc QA				Multi-Doc QA				Dialogue Understanding				Summarization			
	CP	CR	F_1	N	CP	CR	F_1	N	CP	CR	F_1	N	CP	CR	F_1	N
🔒 ***Closed-source LCMs***																
GPT-4o	27.79	**32.17**	28.75	2.02	55.80	**60.65**	55.37	1.71	30.79	**35.70**	32.08	2.17	18.55	25.07	21.00	22.86
Claude-3.5-sonnet	**31.33**	30.20	**30.10**	3.54	**66.05**	56.03	**58.31**	2.44	**36.55**	34.50	**34.90**	2.83	**20.30**	**26.89**	**22.67**	17.70
o1-mini	17.77	20.71	18.68	3.38	44.55	45.10	43.34	1.78	16.75	19.61	17.67	2.63	11.48	16.17	13.13	19.58
🔓 ***Open-source LCMs***																
Qwen2.5-3b-Ins	3.59	3.67	3.47	1.75	14.42	8.72	10.27	1.96	5.13	4.15	4.07	2.33	4.87	5.69	5.04	9.52
Llama3.1-8B-Ins	15.71	17.80	16.23	2.59	41.73	39.45	39.45	1.62	31.43	33.42	31.92	2.08	11.44	12.63	11.86	18.31
Glm4-9B-chat	**20.25**	19.88	19.50	2.21	47.12	43.52	43.83	1.64	28.79	28.19	27.90	1.78	**17.86**	**20.56**	**18.57**	11.38
Qwen2-57B-A14B-Ins	2.08	1.95	1.91	1.27	37.85	23.00	27.45	1.06	11.81	12.16	11.49	1.80	3.87	2.46	2.62	3.81
Llama3.1-70B-Ins	20.20	**21.41**	**20.58**	1.23	**50.11**	**52.92**	**50.59**	1.42	**36.68**	**38.92**	**37.16**	1.76	16.81	19.64	17.87	11.78

Table 12: Citation quality of LCMs within L-CiteEval evaluated by GPT-4o.

E Details of Context Compression and Inference Efficiency Method

E.1 Retrieval-based Method

We utilize the dense retriever GTR-T5-XXL (Ni et al., 2021) to identify the citation chunks that are semantically related to the question. For each question, we select the top 32 citation chunks with the highest retrieval scores and concatenate these segments as input to the LCMs. We conduct experiments on 6 tasks with L-CiteEval benchmark and report the evaluation results in Fig. 5.

E.2 Summarization-based Method

We investigate the use of summarization as a method for context compression. Specifically, we leverage the Meta-Llama-3.1-70B-Instruct model to generate summaries for individual chunks of text. The maximum length of each summary is constrained to be no more than half the length of the original chunk. The summarization process is guided by the prompt: "Summarize the context above concisely in no more than Maximum Tokens tokens."

E.3 Inference Efficiency Method

We report the complete performance of two long-context techniques, StreamingLLM and DuoAttention, on L-CiteEval in Tab. 14. Our findings indicate that when the base model lacks long-context capabilities, long-context efficiency methods do not significantly enhance its performance on long-context tasks. On the other hand, models that are already capable of handling long contexts may suffer from reduced precision in referencing source material when using these efficiency methods, as the acceleration process can potentially discard critical information, leading to less accurate citations in the generated outputs.

F Analysis of Citation Generation

We compare the overall performance of models between those with citation and without citation in

AI assistant’s cited passages: {Model Cited Chunks}
AI assistant’s statement: {Model Generation}
You receive a statement generated by an AI assistant along with passages cited from a document. Your task is to evaluate whether the cited passages adequately support the AI assistant’s statement.
Please follow these guidelines when evaluating:
1. **Rely Only on the Cited Passages**: Base your judgment strictly on the information provided in the cited passages. Do not use any outside knowledge or assumptions.
2. **Ensure Full Coverage**: The cited passages must explicitly and completely support all key details in the statement. If any critical information is missing or ambiguous, the statement should be rated as unsupported.
When providing your evaluation, respond with one of the following ratings:
• **Support**: If the cited passages fully and explicitly support the AI assistant’s statement.
• **Unsupport**: If the cited passages fail to sufficiently support or fully cover the AI assistant’s statement.
Remember: Any missing, unclear, or implied information in the cited passages should result in a rating of **Unsupport**.
Please respond with a single-word rating: ’Support’ or ’Unsupport’ without any explanation.
Your rating:

Figure 4: Prompt for evaluating citation quality with GPT-4o.

Models	**NarrativeQA**		**Natural Questions**		**HotpotQA**		**2WikimultihopQA**		**LoCoMo**		**DialSim**	
	Faith.	**Recall**	**Faith.**	**Recall**	**Faith.**	**Recall**	**Faith.**	**Recall**	**Faith.**	**Recall**	**Faith.**	**Recall**
🔒 ***Closed-source LCMs***												
GPT-4o	60.00	57.50	64.40	**83.20**	87.40	**90.70**	69.50	**76.90**	**88.70**	**73.80**	**77.80**	**70.80**
Claude-3.5-sonnet	**60.10**	**61.20**	**78.90**	83.10	**90.80**	78.40	**82.20**	63.20	75.20	62.80	60.90	53.40
o1-mini	48.00	46.10	70.70	82.30	72.80	72.30	55.40	42.70	59.60	64.70	63.30	57.90
🔓 ***Open-source LCMs***												
Qwen2.5-3b-Ins	20.20	43.80	16.70	71.50	12.00	56.70	17.00	27.50	16.50	52.60	29.00	40.40
Phi-3.5-mini-Ins	36.10	43.30	53.90	76.70	13.20	65.70	8.00	35.90	34.40	64.80	44.30	44.60
Llama-3.1-8B-Ins	44.20	49.90	49.40	80.00	64.20	76.10	50.40	50.40	72.50	66.70	66.90	**60.90**
Glm-4-9B-chat	**47.60**	52.00	**64.70**	**83.80**	72.30	**84.00**	**65.00**	49.90	**80.30**	69.30	59.70	56.20
Mistral-Nemo-Ins	15.60	45.50	13.70	73.40	32.70	70.50	24.30	47.90	40.50	61.20	30.90	57.10
Qwen2-57B-A14B-Ins	15.50	48.30	28.60	79.90	20.70	63.00	12.30	33.60	29.80	55.80	39.80	44.10
Llama-3.1-70B-Ins	46.30	**55.20**	44.10	77.30	**77.50**	79.80	50.30	**58.00**	76.40	**69.50**	**70.00**	55.70
ChatQA-2-70B	20.20	43.50	30.60	75.00	48.90	55.30	19.50	26.30	59.70	55.50	54.80	59.10

Table 13: Faithfulness and Recall of LCMs evaluated with RAGChecker.

Tab. 16. We find that enabling models to generate with citations can remarkably boost the correctness of the model generation in most of the tasks, especially in open-source models. This phenomenon can be attributed to the evidence in Fig.6(b). When models try to generate with citations, they tend to concentrate on the critical chunks.

G Analysis of Attention Mechanism

We explore whether the process of citation generation by LCMs is also reflected in the attention mechanisms. Let the ground truth citation segment within the context be denoted as g_j. Following Wu et al. (2024), we can use the retrieval score to determine whether the LCM’s attention focuses on the segment containing g_j when generating the citation for g_j. We find the positions that receive the most attention from all the attention heads. If a position is located in the segment containing g_i and the model’s output citation is exactly g_i, or if neither matches, we consider this a “correct retrieval”. Otherwise, it is an “incorrect retrieval”. We conduct the experiments on two tasks (HotpotQA and 2WikiMultihopQA) with two strong LCMs (Llama-3.1-8B-Instruct and GLM-4-9B-Chat). We plot the number of citations generated by the models and the number of citation segments identified by the attention heads in Fig. 6(a). We utilized Pearson correlation analysis to calculate the correlation coefficient (r) between the generated citations and those retrieved by the attention mechanism, finding all the correlation values exceed 0.7. This reveals

Models	**NarrativeQA**		**HotpotQA**		**GovReport**		**LoCoMo**		**Counting Stars**	
	F_1	**Rec.**	**F_1**	**Rec.**	**F_1**	**Rou.**	**F_1**	**Rec.**	**F_1**	**Acc**
Llama-2-7b-chat	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
+ StreamingLLM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Llama-3-8B-Ins-Gradient	20.17	63.76	4.33	67.81	0.83	25.44	11.88	61.51	1.87	5.00
+ DuoAttention	7.95	61.96	2.58	70.31	0.00	25.43	7.41	47.98	3.47	27.50

Table 14: Citation results and generation results of long-context techniques where **F_1** denotes citation quality, **Rec.** denotes recall score and **Rou.** denotes Rouge-L score.

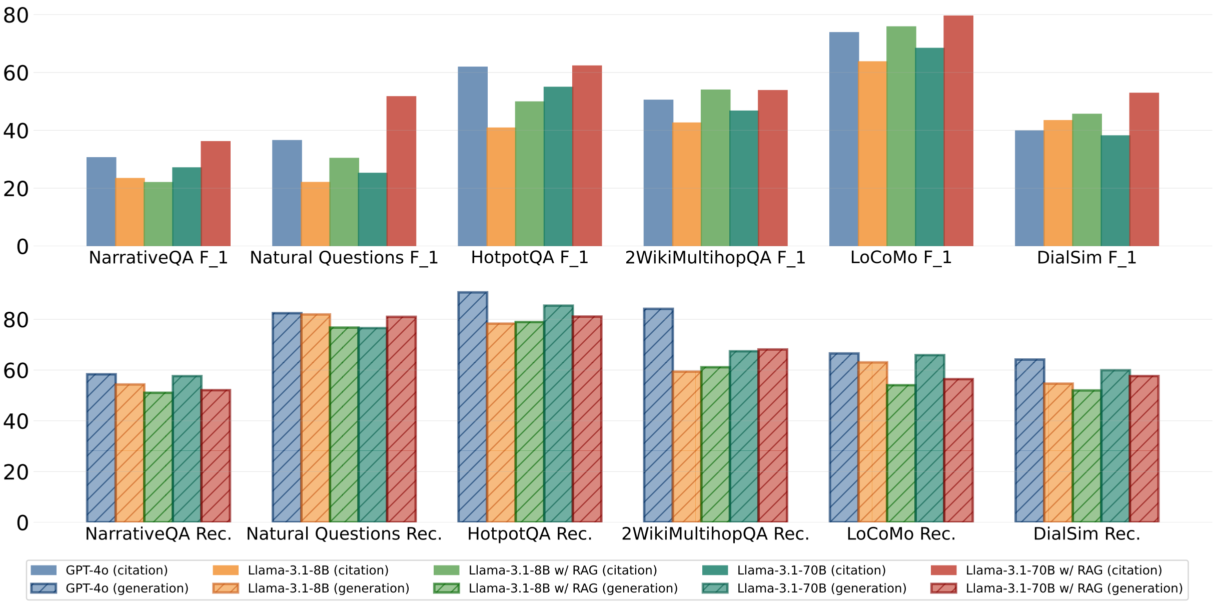


Figure 5: Performance of RAG on 6 tasks in L-CiteEval, where the top group shows citation quality and the bottom group shows generation quality.

the underlying mechanism by which we can leverage the model's citation output to verify whether the model is truly responding based on the given context.

We also calculate the recall rate of the top 10 positions where models focus within the golden segments across three datasets. The results are presented in Fig.6(b). The findings suggest that generating with citations allows models to identify evidence related to the answer more effectively compared with directly generating.

H Analysis on model intrinsic knowledge

We conduct experiments to investigate the phenomenon whereby models tend to rely on their internal knowledge rather than basing their responses solely on the provided context. We utilize the counterfact dataset for evaluation. First, we identify which factual knowledge the model inherently possesses. Then, we insert the corresponding counterfactual information as the needle into a long context to test the NIAH task. The results confirm our hypothesis: even when the model cites the correct passage, it may still respond based on its own knowledge rather than the provided information. Two illustrative cases are presented in Tab. 17.

I Comparison between L-CiteEval and other Long-Context Benchmarks

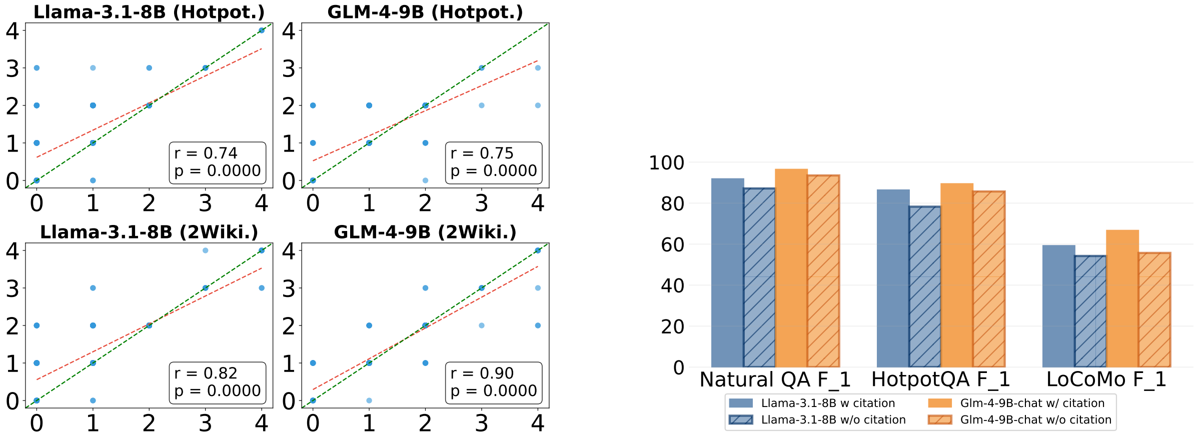
We present specific results to compare L-CiteEval with LongBench and Ruler in Fig. 7. L-CiteEval assesses LCMs from two unique perspectives: citation quality and generation quality, thereby enhancing the distinctions in performance among LCMs.

J Cases study

We provide several model generation results from Fig. 8 to Fig. 11.

Models	*L-CiteEval-Length*						*L-CiteEval-Hardness*					
	0∼8k		**8∼16k**		**16∼32k**		**Easy**		**Medium**		**Hard**	
	`Cite`	`Res.`	`Cite`	`Res.`	`Cite`	`Res.`	`Cite`	`Res.`	`Cite`	`Res.`	`Cite`	`Res.`
						NarrativeQA						
GPT-4o-2024-05-13	62.08	62.63	46.67	61.36	33.25	64.84	40.83	100.00	46.25	69.67	54.92	19.16
Qwen2.5-3b-Ins	17.50	56.19	4.58	58.09	1.25	56.96	11.67	75.00	4.58	60.02	7.08	36.22
Llama-3.1-8B-Ins	43.01	61.99	39.17	64.41	40.27	62.55	27.92	94.17	52.08	69.78	42.44	25.00
Qwen2-57B-A14B-Ins	12.50	58.52	0.00	51.12	12.92	53.41	5.00	75.00	15.42	63.13	5.00	24.92
Llama-3.1-70B-Ins	59.17	63.42	51.67	63.24	47.50	62.86	43.75	94.17	55.83	70.76	58.75	24.60
						HotpotQA						
GPT-4o-2024-05-13	65.67	95.67	63.50	95.25	63.75	89.62	66.50	100.00	71.42	100.00	55.00	80.54
Qwen2.5-3b-Ins	3.81	70.42	6.58	65.21	4.76	55.62	3.81	71.25	3.67	66.46	7.68	53.54
Llama-3.1-8B-Ins	51.83	85.25	40.56	81.04	40.83	67.75	52.67	88.75	41.39	82.29	39.17	63.00
Qwen2-57B-A14B-Ins	12.50	85.62	7.29	72.92	6.83	62.92	12.50	83.12	5.62	73.33	8.50	65.00
Llama-3.1-70B-Ins	67.50	89.42	56.67	90.25	49.17	85.25	65.83	91.25	59.17	92.50	48.33	81.17
						GovReport						
GPT-4o-2024-05-13	56.68	23.07	48.82	21.48	44.45	20.65	49.95	15.26	51.27	10.86	48.74	9.24
Qwen2.5-3b-Ins	21.12	27.66	13.08	28.16	3.43	22.92	14.32	16.28	9.31	14.65	14.00	14.37
Llama-3.1-8B-Ins	57.08	24.27	38.28	24.15	18.46	19.25	40.35	15.55	34.75	12.09	38.72	12.57
Qwen2-57B-A14B-Ins	6.55	29.51	2.09	30.52	1.71	24.20	3.48	30.02	3.26	25.37	3.61	28.85
Llama-3.1-70B-Ins	57.55	25.41	43.60	23.43	17.64	21.62	37.47	16.36	35.46	13.04	45.86	11.98
						LoCoMo						
GPT-4o-2024-05-13	78.13	68.07	73.91	66.93	72.24	68.77	78.52	100.00	71.37	85.30	74.39	18.47
Qwen2.5-3b-Ins	16.40	55.18	10.81	45.12	6.77	43.87	8.44	69.12	15.85	60.09	9.70	14.96
Llama-3.1-8B-Ins	76.51	68.68	63.54	68.39	63.91	61.33	76.17	96.62	70.07	82.06	57.72	19.73
Qwen2-57B-A14B-Ins	55.92	63.76	22.92	58.18	16.13	59.29	44.17	84.23	15.58	73.67	35.21	23.32
Llama-3.1-70B-Ins	75.45	73.21	71.27	70.53	64.38	57.89	81.64	93.56	67.24	79.3	62.21	28.76
						Counting Stars						
GPT-4o-2024-05-13	97.30	93.33	92.71	83.33	92.95	88.75	100.00	100.00	100.00	100.00	82.96	65.42
Qwen2.5-3b-Ins	2.67	37.08	5.17	32.50	0.00	29.58	1.33	36.67	4.51	40.00	2.00	22.50
Llama-3.1-8B-Ins	42.93	42.08	35.64	33.75	18.70	20.00	40.18	32.50	30.05	28.33	27.04	35.00
Qwen2-57B-A14B-Ins	27.21	45.00	10.51	77.92	0.89	46.25	21.71	49.17	5.74	57.08	11.16	62.92
Llama-3.1-70B-Ins	76.96	56.67	74.93	66.25	65.14	58.33	77.16	54.17	69.21	58.75	70.66	68.33

Table 15: Model performance on *L-CiteEval-Length* and *L-CiteEval-Hardness*, where we report `F_1` score to reflect citation quality (`Cite`) and recall/rouge-L/accuracy for different downstream tasks to reflect generation quality (`Res.`).



(a) (b)

Figure 6: Analysis of attention mechanism: (a) Pearson correlation analysis between generated citations and attention mechanisms. The x-axis represents the number of correct citations produced by the model, and the y-axis represents the number of correct citation segments attended by the attention. The red curve indicates the fitted correlation, with closer alignment to the green curve signifying a higher correlation. (b) The recall rate of the top 10 positions where models focus in the golden segments.

Model	**Single-Doc QA**		**Multi-Doc QA**		**Summ.**	**Dialogue**		**Synthetic**	
	Prec.	**Rec.**	**Prec.**	**Rec.**	**Rouge-L**	**Prec.**	**Rec.**	**Rouge-1**	**Acc**
GPT4o	11.78	70.37	10.34	**87.38**	**20.15**	9.81	**65.35**	89.24	**91.88**
w/o citation	**12.18**	**70.59**	**11.09**	85.09	19.00	**10.29**	64.93	**90.62**	90.42
LLama3.1-8B-Ins	10.11	**68.13**	**7.66**	**68.84**	**20.90**	**11.07**	58.84	**85.11**	**33.75**
w/o citation	**10.56**	64.83	4.46	55.00	18.40	7.61	**59.09**	61.63	12.22
GLM4-9B-chat	**11.22**	**67.25**	**7.88**	**77.97**	**21.42**	**7.69**	51.25	**90.81**	**58.82**
w/o citation	8.27	66.85	6.55	71.25	18.35	5.42	**54.93**	87.33	38.68

Table 16: Comparisons of model generations in terms of correctness between w/ citation and w/o citation.

Case 1	
Context	[...]Passage [33]: [...]Chevrolet Brookwood is sold by Intel. [...]Passage [34]:[...]Write an accurate, engaging, and concise answer to the given question using only the provided passages (some of which might be irrelevant). Use an unbiased and journalistic tone. Every sentence must include a citation at the end, referencing at least one passage and at most three. When citing several passages, use separate brackets for each index number, like [a][b][c], instead of combining them in one set of brackets, like [a, b, c]. Here, a, b and c represent different index numbers. If multiple passages support the sentence, only cite a minimum sufficient subset of the passages. \n\n Question: Who sells the Chevrolet Brookwood?\nAnswer:
Question	Who sells the Chevrolet Brookwood?
Model Generation	Chevrolet sells the Brookwood[34].
Counterfact Answer	Intel
Fact Answer	Chevrolet
Case 2	
Context	[...]Passage [33]: [...]Paris is a part of the continent of Antarctica. [...]Write an accurate, engaging, and concise answer to the given question using only the provided passages (some of which might be irrelevant). Use an unbiased and journalistic tone. Every sentence must include a citation at the end, referencing at least one passage and at most three. When citing several passages, use separate brackets for each index number, like [a][b][c], instead of combining them in one set of brackets, like [a, b, c]. Here, a, b and c represent different index numbers. If multiple passages support the sentence, only cite a minimum sufficient subset of the passages. \n\n Question: What continent can one navigate to get to Paris? \nAnswer:
Question	What continent can one navigate to get to Paris?
Model Generation	Paris is a part of the continent of Europe[33].
Counterfact Answer	Antarctica
Fact Answer	Europe

Table 17: Cases where the output of Llama-3.1-8B-Instruct is based on its inherent knowledge rather than the provided context.

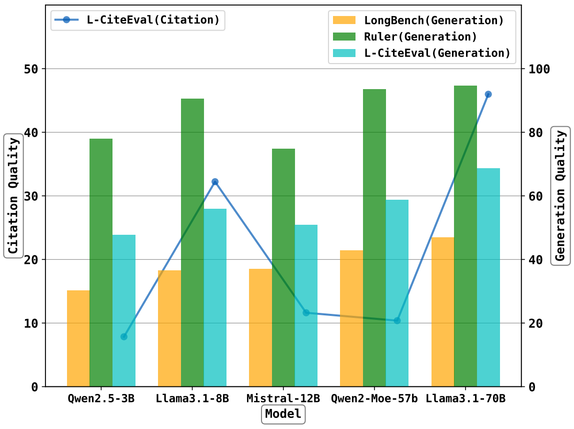


Figure 7: Comparison between L-CiteEval and other long-context benchmarks.

Llama-3.1-70B-Instruct in LoCoMo

Context:

...

Passage [13] :

```
... Nate said, \"Coconut milk ice cream is at the top of my list.
It's so smooth and creamy with a tropical coconut twist. Plus,
it's dairy-free for people who can't have lactose or who  ...
```

...

Question:

What is Nate's favorite dish from the cooking show he hosted?

Answer:

Nate's favorite dish from the cooking show he hosted is coconut milk ice cream [13].

Ground truth:

Coconut milk ice cream

Figure 8: Case study of Llama-3.1-70B-Instruct on LoCoMo task.

GPT-4o in NarrativeQA

Context:

...

Passage [44]:

```
On the men's faces in strained anxiety. <b> ... you are to disa-
rm yourself and return immediately to the bomb bay.\n Do you un-
derstand? <b>  BOMB \#20\n</b>  (over)\n I am programmed to det-
onate in\n  fourteen minutes thirty seconds.\nDetonation will
occur at the\n  programmed time...
```

Passage [45]:

```
BOMB \#20\n</b>  I read you, Lieutenant Doolittle,\n ut I am pr-
ogrammed to detonate in\n  fourteen minutes.  Detonation will\n
occur at the programmed time...
```

Passage [49]:

```
...PINBACK\n</b> But you can't explode in the bomb\n bay. It
's foolish. You'll kill us\n all.  There's no reason for it. <b>
BOMB \#20\n</b> (over)\n  I am programmed to detonate in nine
```

Passage [53]:

```
All right now, here's the big\n  question: how do you know that
the\n evidence your sensory apparatus\n  reveals to you is corr-
ect? ... Talby lies unconscious near the burned laser.
```

...

Question:

What self-detonates despite several attempts to stop it?

Answer:

Bomb #20 self-detonates despite several attempts to stop it [44][45][49][53].

Ground truth:

Bomb #20.

Figure 9: Case study of GPT-4o on NarrativeQA task.

Claude-3.5-sonnet in HotpotQA

Context:

...

Passage [39] :

`Gerd Neggo (9 November 1891 \u2013 1 September 1974) was an Estonian dancer, dance teacher and choreographer. She studied the musical response methods of \u00c9. Jaques-Dalcroze, trained under Rudolf von Laban in Hamburg, Germany, and in 1924 established`

Passage [49] :

`Rudolf von Laban, also known as Rudolf Laban ... (15 December 1879 \u2013 1 July 1958), was a dance artist and theorist. He is notable as one of the pioneers of modern dance in Europe. His work laid the foundations ...`

...

Question:

Gerd Neggo trained under the founder of which type of dance analysis?

Answer:

Gerd Neggo trained under Rudolf von Laban in Hamburg, Germany [39]. Rudolf von Laban is notable as one of the pioneers of modern dance in Europe and his work laid the foundations for Laban Movement Analysis [49]

Ground truth:

Laban Movement Analysis

Figure 10: Case study of Claude-3.5-sonnet on HotpotQA task.

Llama-3.1-8B-Instruct in MultiNews

Context:

...

Passage [1]:

The Pennsylvania State Police have a rigorous application process, which includes a polygraph pretest. Recently, a Crawford County man applying to become a state trooper apparently failed his polygraph so spectacularly ... Only that White allegedly told the examiner that he had had consensual sex and other contact with the girl in 2011...

Passage [2]:

She is now 19. Police tell AP they charged White on Thursday with ... Needless to say, he won't be getting an interview.

Passage [3]:

MEADVILLE, Pa. (AP) \u2014 Police say a Pennsylvania man applying to become a state trooper has been arrested after saying during a polygraph pretest that he had sex with an underage girl four years ago... She is now 19. Police say they charged White on Thursday with four counts of unlawful sexual contact with a minor and 10 counts of corruption of minors after interviewing the woman and corroborating that information.

Passage [4]:

and he doesn't have a listed phone number...

...

Summary:

A Pennsylvania man, Joseph Adam White, was arrested after he failed a polygraph test during his application to become a state trooper. During the test, he admitted to having sex with an underage girl four years ago [1][2]. The girl is now 19 years old [2]. White was charged with four counts of unlawful sexual contact with a minor and 10 counts of corruption of minors [2]. He was arrested at the Meadville barracks where he was taking the polygraph test [3].

Ground truth:

2013 A Pennsylvania man is behind bars after allegedly telling state police he'd had sex with an underage girl. What's unusual: He told them while applying for a job as a state trooper, the AP reports via the Patriot-News. Police say Joseph White, 29, made the admission during a polygraph pretest at the Meadville state police barracks on Jan. 15. He allegedly admitted to consensual sex and other sexual contact with the girl four years ago; she's now 19. Police say the victim later corroborated the incidents, which occurred in a field in South Shenango Township, the Meadville Tribune reports. White was arrested Thursday on multiple charges of unlawful sexual contact with a minor and corruption of minors, and jailed on $25,000 bond. N̈eedless to say, he won't be getting an interview,q̈uips the Patriot-News.

Figure 11: Case study of Llama-3.1-8B-Instruct MultiNews task.