L-CITEEVAL: DO LONG-CONTEXT MODELS TRULY LEVERAGE CONTEXT FOR RESPONDING?

Zecheng Tang[1], **Keyan Zhou**[1]*, **Juntao Li**[1]†, **Baibei Ji**[1], **Jianye Hou**[2], **Min Zhang**[1]
[1]Soochow University [2]CUHK
{zctang, kyzhou, bbjidbj}@stu.suda.edu.cn,
{ljt, minzhang}@suda.edu.cn, jianyehou@link.cuhk.edu.cn

Code & Data: https://github.com/ZetangForward/L-CITEEVAL.git

ABSTRACT

Long-context models (LCMs) have made remarkable strides in recent years, offering users great convenience for handling tasks that involve long context, such as document summarization. As the community increasingly prioritizes the faithfulness of generated results, merely ensuring the accuracy of LCM outputs is insufficient, as it is quite challenging for humans to verify the results from the extremely lengthy context. Yet, although some efforts have been made to assess whether LCMs respond truly based on the context, these works either are limited to specific tasks or heavily rely on external evaluation resources like GPT-4. In this work, we introduce ***L-CiteEval***, a comprehensive multi-task benchmark for long-context understanding with citations, aiming to evaluate both the understanding capability and faithfulness of LCMs. L-CiteEval covers 11 tasks from diverse domains, spanning context lengths from 8K to 48K, and provides a fully automated evaluation suite. Through testing with 11 cutting-edge closed-source and open-source LCMs, we find that although these models show minor differences in their generated results, open-source models substantially trail behind their closed-source counterparts in terms of citation accuracy and recall. This suggests that current open-source LCMs are prone to responding based on their inherent knowledge rather than the given context, posing a significant risk to the user experience in practical applications. We also evaluate the RAG approach and observe that RAG can significantly improve the faithfulness of LCMs, albeit with a slight decrease in the generation quality. Furthermore, we discover a correlation between the attention mechanisms of LCMs and the citation generation process.

1 INTRODUCTION

The rapid development of Long-context Models (LCMs) provides users with numerous conveniences in resolving long-context real-world tasks, such as code analysis (Zhu et al., 2024) and long document summarization (Reid et al., 2024). Recently, the community has gradually intensified its efforts to enhance the faithfulness of generative artificial intelligence (Manna & Sett, 2024), which is of paramount importance for LCMs. This is because tasks that involve long context usually require LCMs to respond based on the provided context rather than relying solely on models' intrinsic knowledge. Therefore, there is an urgent need for a benchmark to verify whether LCMs truly leverage context for responding and reflect those models' capability on long-context tasks.

To date, substantial efforts have been made to develop benchmarks for evaluating LCMs. These endeavors aim to achieve several key objectives: (1) ensuring that the benchmarks include a **comprehensive** range of task scenarios and varying context lengths; (2) employing automated metrics to guarantee the **reproducibility** of evaluations; (3) incorporating an appropriate volume of test data to maintain evaluation **efficiency**; and (4) offering sufficient **interpretability** (e.g., providing

*Equal Contribution
†Corresponding Author

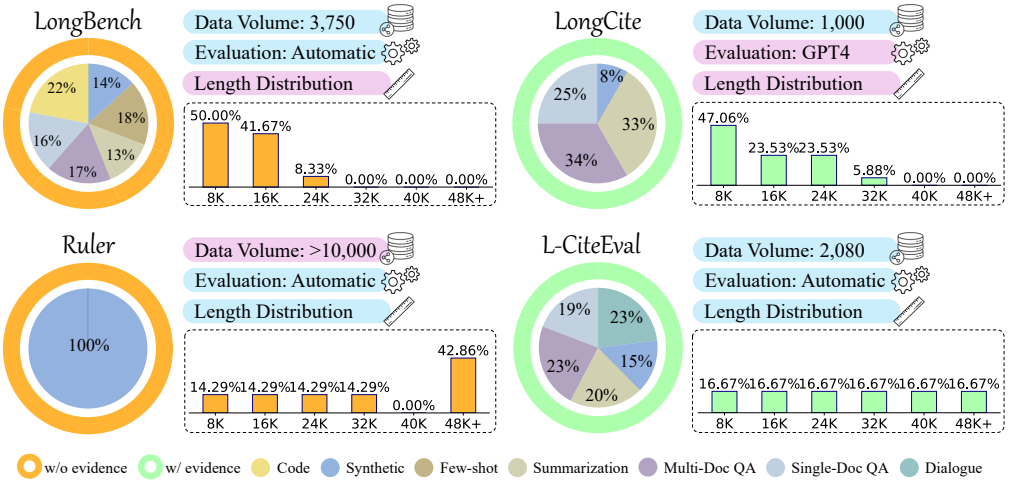

Figure 1: Overview and comparison among different representative benchmarks for LCMs.

evidence to support the responses). As shown in Fig. 1, taking three representative long-context benchmarks as examples: LongBench (Bai et al., 2023) primarily evaluates the accuracy of LCMs' responses across a range of realistic and synthetic tasks, with a context length of up to 24K tokens; Ruler (Hsieh et al., 2024) focuses on using synthetic data to test LCMs' capabilities in information retrieval over long sequences, with context lengths exceeding 48K tokens; and LongCite (Bai et al., 2024) assesses whether models respond based on the content within the context, employing GPT-4 as a judge. These benchmarks, based on their purpose, can be roughly divided into two categories: (1) evaluating long-context understanding and (2) assessing model faithfulness. The former evaluates model outputs using large volumes of test data to infer LCMs' capabilities but lacks interpretability to the generated results. The latter are mainly based on short-context datasets (e.g., in LongCite, the maximum sequence length only reaches 32K, comprising just 5.88% of the benchmark) and rely on external resources like GPT-4 to judge faithfulness, making the evaluation results hard to reproduce. In this work, we introduce ***L-CiteEval***, a comprehensive multi-task benchmark for long-context understanding with citations. As shown in Fig. 2, given the question and long reference context, L-CiteEval requires LCMs to generate both the statements and their supporting evidence (citations). There are **5** major task categories, **11** different long-context tasks, with context lengths ranging from **8K** to **48K** in L-CiteEval. To address the timeliness and the risk of data leakage in testing (Ni et al., 2024; Apicella et al., 2024), we incorporate 4 latest long-context tasks as the subsets in L-CiteEval, ensuring that the evaluation remains up-to-date and robust. Different from previous benchmarks for long-context understanding that primarily assess LCMs based on their predicted answers, L-CiteEval evaluates model performance based on both the generation quality (whether the predicted answer is correct) and citation quality (whether the provided citations can support the corresponding answer). To extend the context length of short-context data, we design a rigorous data construction pipeline to extend the sequence length and mitigate the perturbation introduced from the additional context. Additionally, to facilitate the ease of use and ensure reproducibility, L-CiteEval offers an automatic evaluation suite. Considering that the prediction from LCMs can be influenced by both the task difficulty and the context length, we propose two benchmark variants: ***L-CiteEval-Length*** and ***L-CiteEval-Hardness***. These two variants strictly control the variables within the evaluation, focusing solely on context length and task difficulty to assess LCMs' capabilities.

We test 11 cutting-edge and widely-used LCMs, including 3 closed-source and 8 open-source models, which feature different sizes and architectures. We also explore whether the Retrieval-Augmented Generation (RAG) technique can improve the faithfulness of LCMs. Evaluation results indicate that there is a minor difference between open-source and closed-source models regarding generation quality, while open-source models substantially trail behind their closed-source counterparts in terms of citation quality. Utilizing the RAG technique exhibits a notable improvement in the faithfulness of open-source models, but it slightly impacts the generation quality. Furthermore, we reveal a correlation between the model's citation generation process and its attention mechanism (i.e., retrieval head (Wu et al., 2024)), demonstrating the validity of our benchmark and offering insights for future evaluations of LCM faithfulness and the development of advanced LCMs.