
LaRA: Benchmarking Retrieval-Augmented Generation and Long-Context LLMs — No Silver Bullet for LC or RAG Routing

Kuan Li¹ Liwen Zhang² Yong Jiang² Pengjun Xie² Fei Huang² Shuai Wang¹ Minhao Cheng³

Abstract

As Large Language Model (LLM) context windows expand, the necessity of Retrieval-Augmented Generation (RAG) for integrating external knowledge is debated. Existing RAG vs. long-context (LC) LLM comparisons are often inconclusive due to benchmark limitations. We introduce LaRA, a novel benchmark with 2326 test cases across four QA tasks and three long context types, for rigorous evaluation. Our analysis of eleven LLMs reveals the optimal choice between RAG and LC depends on a complex interplay of model capabilities, context length, task type, and retrieval characteristics, offering actionable guidelines for practitioners. Our code and dataset is provided at: <https://github.com/Alibaba-NLP/LaRA>.

1. Introduction

While large language models (LLMs) excel across various domains, the dynamic nature of information poses significant challenges to their ability to acquire new knowledge effectively. Current studies reveal several limitations of LLMs, including high computational costs when processing long texts, a tendency to produce factual errors and hallucinations, difficulty adapting to specialized domains, and a propensity for generating overly generic responses (Gao et al., 2023). To address these limitations, researchers have explored retrieval-augmented generation (RAG) (Gua et al., 2020; Lewis et al., 2020). RAG enables LLMs to efficiently utilize external knowledge by retrieving the most relevant fragments from uploaded

documents, knowledge bases, or websites. However, recent advancements in LLMs, such as GPT-4o (OpenAI, 2024), Llama 3.2 (Meta, 2024a), Claude 3.5 (Anthropic, 2024), and Qwen 2.5 (Yang et al., 2024), now support input lengths of up to 128k tokens, offering an alternative by directly feeding the full context of relevant information into the model. This raises questions about the continued necessity of RAG, which was initially crucial for handling long texts, since these models can now potentially access and process the necessary information directly. Therefore, it is essential to systematically compare the strengths and weaknesses of RAG and long-context (LC) LLMs.

Numerous studies have investigated the performance differences between RAG and providing LLMs with full long contexts. For example, Xu et al. find that RAG outperforms LC on several traditional QA datasets. More recently, Li et al. argued that LC consistently outperforms RAG in almost all settings. However, Yu et al. subsequently conducted experiments demonstrating that RAG is not inherently weaker than LC. This lack of consensus likely stems from several flaws in the evaluation pipeline design of existing benchmarks, including issues with the corpus (e.g., excessively short text lengths, failed replacements), evaluation metrics, and impractical task design.

To address these issues and facilitate a robust comparison between RAG and LC, we propose LaRA, a benchmark for evaluating Long-context LLMs competing against RAG. In constructing LaRA, we adhere to the following criteria: (1) Context length should be maximized within the LLMs’ input limits to avoid truncation that could obscure the true capabilities of the models. (2) The context should consist of naturally occurring long text, rather than artificially constructed examples, to reflect real-world usage scenarios. (3) Answering questions should require information from the provided context, ensuring the LLM cannot answer them based solely on its internal knowledge. (4) Questions should have definitive answers to facilitate an accurate evaluation using LLMs. (5) The questions should reflect practical queries that humans are likely to ask in real-world LLM usage scenarios. LaRA is designed to serve as a guidebook for designing effective RAG-or-LC planning systems, therefore, it is crucial that all included questions are meaningful,

The project was done during Kuan Li’s internship at Tongyi Lab, Alibaba Group. ¹The Hong Kong University of Science and Technology ²Tongyi Lab, Alibaba Group ³Penn State University. Correspondence to: Yong Jiang <yongjiang.jy@alibaba-inc.com>, Shuai Wang <shuaiw@cse.ust.hk>, Minhao Cheng <mmc7149@psu.edu>.

relevant, and reflect real-world information needs.

Specifically, LaRA includes three of the most common types of long contexts—novels, academic papers, and financial statements—selected to represent diverse writing styles and information densities. LaRA features four types of question-answering tasks designed to assess essential capabilities for handling long contexts: Locating specific information, Comparing different parts of the text, Reasoning about the content, and Detecting hallucinations. The QA pairs are constructed using a combination of human annotation and LLM-assisted generation, where we iteratively refine seed examples and prompts until a predefined pass rate is achieved for the generated pairs. To ensure accurate and reliable evaluation, we employ GPT-4o as a judge to determine the correctness of predictions and further validate the results by computing Cohen’s Kappa coefficient between LLM and human evaluations, confirming a high level of consistency.

Our extensive experiments on LaRA demonstrate that the choice between RAG and LC is not trivial, as it varies significantly depending on factors such as model size, query type, type of tasks, context length, context type, and number of retrieved chunks. If we can pinpoint the scenarios in which RAG outperforms LC, we can better design workflows to route each query through RAG or LC, thereby optimizing for both cost and performance, leading to more efficient and effective LLM applications. Our key findings are as follows:

- **Model Strength:** RAG provides more significant improvements for weaker models. Our analysis indicates a correlation between model strength and RAG’s effectiveness: the weaker the model, the greater the improvement from RAG. For instance, with a 128k context length, RAG outperformed LC by 6.48% and 38.12% in accuracy on Llama-3.2-3B-Instruct and Mistral-Nemo-12B, respectively. For models with strong long-text capabilities, such as GPT-4o and Claude-3.5-sonnet, LC generally outperforms RAG, demonstrating the effectiveness of these models in directly processing extensive contexts.
- **Context Length:** RAG’s advantages become more pronounced as context length increases. With a 32k context length, LC achieved an average accuracy 2.4% higher than RAG across all models. However, with a 128k context length, this trend reversed, with RAG outperforming LC by 3.68%.
- **Task Performance:** RAG demonstrates similar performance to LC in single-location tasks and offers a significant advantage in identifying hallucinations. In contrast, LC excels in reasoning tasks and comparison tasks.

2. Revisiting RAG vs. Long Context Benchmarks

Despite a lot of benchmarks has been used to compare RAG with feeding language model with long context, there lacks a clear guidelines and conclusion on when and where the RAG will be a better choice than long context. [Xu et al.](#) and [Bai et al.](#) draw opposing conclusions on whether RAG or LC performs better on traditional QA datasets. Recently, [Li et al.](#) argued that LC consistently outperforms RAG in almost all settings, and [Yu et al.](#) subsequently claim RAG can defeat LC on the same benchmark. In this section, we conduct a detailed analysis on the existing benchmark and analysis, and find the key issues are stem from some significant flaws with their evaluation pipeline. For simplicity, our analysis is mainly limited to question answering tasks based on long contexts.

2.1. Issues with External Contexts

Insufficient context lengths. As LLM base models continue to evolve, the definition of “long context” has also shifted, expanding from the early limit of 4k tokens to the now commonly supported 128k context length. Some early work utilize datasets such as Qasper (QASP) ([Dasigi et al., 2021](#)), NarrativeQA (NQA) ([Kociský et al., 2018](#)), and QuALITY (QLTY) ([Pang et al., 2022](#)) to compare RAG and LC. For instance, [Xu et al.](#) conduct experiments on these datasets and find that RAG can strengthen large models, such as Llama2-70B and GPT-43B. Similarly, [Lee et al.](#) combine these datasets to create a new benchmark for further evaluations. However, such datasets no longer align with the current definition of long context. For example, QASP and QLTY have average context lengths of only 4912 and 6592 tokens, respectively, which are far below the context length capabilities of modern LLMs. Moreover, RAG typically uses chunk sizes of 300–600 tokens, and with 5–20 retrieved chunks, the total context length in RAG becomes comparable to that of full-context input, reducing the distinction between the two approaches.

Data Leakage. Since LLMs use more and more datasets in the training procedure, the problem of data leakage becomes more serious. At the same time, it is challenging to verify whether these early datasets were part of the training data for LLMs, potentially causing the models to memorize the answers. For example, although NarrativeQA has an average context length of 84,770 tokens, Gemini 1.5pro achieves 100% accuracy on this dataset ([Lee et al., 2024](#)), indicating that either the dataset itself or its contexts were likely included in the model’s training process.

Inappropriate Contexts Handling. ∞ -bench ([Zhang](#)

Table 1. Results on En.MC, EN.QA, and Zh.QA in ∞ -bench.

Task	GPT-4o	Qwen 2.5-7B	Qwen 2.5-7B (vote)
En.MC	76.0	67.7	87.9
En.QA	31.5	20.3	31.2
En.QA (LLM)	82.9	78.1	85.5
Zh.QA	26.3	18.4	25.6
Zh.QA (LLM)	79.9	72.5	85.2

et al., 2024a), a recent benchmark widely used for comparing RAG and LC (Li et al., 2024b; Yu et al., 2024), includes two tasks, En.QA and En.MC, with average context lengths of 192.6k and 184.4k tokens, respectively. However, their method of handling the excessive context length is problematic. When the context exceeds the LLM’s maximum context length, the middle portion of the context is truncated. Given that most models have a limit of 128k tokens or even less, it is highly likely that the answer to a question is removed during truncation. In such cases, failure to answer reflects the truncation issue rather than the model’s capabilities. To verify this, we split the excessively long cases in En.QA and En.MC into several parts, each within the context length limit. These parts are then paired with the query and fed into the LLM separately. The LLM is instructed to first determine whether the question is answerable; if not, it would decline to provide an answer. After obtaining multiple answers, a voting mechanism is used to get the final answer. The results are list in Table 1. We observe that after adopting the segmented input and voting mechanism, Qwen-2.5-7B can even outperform GPT-4o.

Moreover, to prevent overlap with data seen during LLM training, key entity replacement is employed as a countermeasure in ∞ -bench. However, upon closer inspection, we find that some replacements are unsuccessful. For example, some entities mentioned in the questions do not exist in the provided context and vice versa¹.

2.2. Inaccurate Evaluation

Unreasonable metrics. Many previous evaluations use automated metrics such as F1-score and exact match (EM) (Chen et al., 2024a; Zhang et al., 2024c), which are not reliable for NLG (Novikova et al., 2017). For example, if the ground truth answer is “Allyson Kalia” and the model’s response is “Allyson Kalia is convicted of the murder of Kiran’s younger brother, Rosetta.” the prediction is clearly correct. However, it would only achieve an F1-score of 0.29. This is also why the scores on the En.QA task in ∞ -bench (Zhang et al., 2024a) tend to be very low. We use LLM to re-evaluate, and the results are shown in Table 1. The accuracy becomes much higher, indicating that these

¹<https://github.com/OpenBMB/InfiniteBench/issues/26>

datasets are not as difficult as they appear.

No dedicated benchmarks. Over the past year, several long-context benchmarks have been introduced, e.g., ZeroSCROLLS (Shaham et al., 2023), LongBench (Bai et al., 2024a), BAMBOO (Dong et al., 2024), LooGLE (Li et al., 2024a), Ruler (Hsieh et al., 2024), ∞ -bench (Zhang et al., 2024a), and LongBench-V2 (Bai et al., 2024b), with text lengths progressively increasing from 20k to 200k tokens. However, these benchmarks focus primarily on testing the ability of models to handle long texts. Although some of these benchmarks include experiments related to RAG, they lack a more systematic comparison between RAG and LC. Furthermore, existing RAG benchmarks typically utilize context lengths under 10k tokens (Chen et al., 2024a; Zhang et al., 2024c; Stolfo, 2024; Lyu et al., 2024; Saad-Falcon et al., 2024), failing to adequately address the challenges of long contexts. Although concurrent work like LONG²RAG (Qi et al., 2024) explores long-context RAG, its average text length remains smaller than LaRA, and its focus lies on evaluating long-form responses. Similarly, Loong (Wang et al., 2024a) introduces tasks for comparing RAG and LC in multi-document QA but suffers from homogeneity in queries. For example, the clustering task only involves determining citation relationships between papers, and reasoning questions merely require providing citation chains. These queries, disconnected from specific content and applicable to any similar text, fail to capture the generalization capabilities of LLMs and RAG across diverse scenarios and context distributions.

3. LaRA

In this section, we introduce the construction of LaRA and how it addresses the issues present in previous benchmarks, as mentioned in Section 3. The statistics of LaRA are provided in Appendix A.

3.1. Long Context Data Collection

In our context selection process, we adhere to the following principles: (1) Timeliness: We select **recent** high-quality long contexts to prevent data leakage issues, ensuring that they are less likely to have been included in the LLM’s training data. (2) Appropriate Length: Considering that mainstream commercial and open-weight models typically support context length of 32k and 128k, we choose contexts that are as close to these window sizes as possible without exceeding them. (3) Naturalness: The chosen contexts are naturally occurring long documents, rather than artificially constructed or assembled from unrelated short texts, to ensure the benchmark reflects the complexity and diversity of real-world use. (4) Authoritativeness: All contexts are considered reliable and credible sources of information due

to expertise, reputation, and qualifications of the authors or institutions behind them.

To ensure a diverse range of contexts, we select novels², financial statements³, and academic papers⁴ as the context. For novels, we choose the txt format of novelettes and novels to serve as the 32k and 128k contexts, respectively. Financial statements include the latest quarterly reports (32k) and annual reports (128k) from publicly listed companies in the United States for the year 2024. To create contexts of appropriate length for academic papers, we combine several papers published on arXiv in 2024 that are related through citations.

Entity Replacement. To mitigate the risk of data leakage from novels, which are likely present in LLMs’ training data, we perform entity replacement. Previous work has employed similar strategies (Zhang et al., 2024a; Li et al., 2022), but we find that many entity replacements were incorrect or inconsistent, leading to inaccurate evaluations. To address this, we use GPT-4o to accurately identify and replace character entities as well as formulating questions targeting the replaced entities, ensuring consistency between the novel text and the questions. Details are provided in Appendix B.

3.2. Tasks in LaRA

To comprehensively evaluate the capabilities of LC LLMs and RAG, LaRA includes four major task categories: location, reasoning, comparison, and hallucination detection, which are designed to assess distinct aspects of LLM performance, motivated by the need to assess both the strengths and weaknesses of RAG and LC in handling complex, real-world information needs. Below, we will introduce each task in detail and further elaborate on the motivation behind them. Examples of each task are provided in Appendix F.

Location. The location task, the most fundamental task in LaRA, evaluates an LLM’s ability to locate specific information within a long context. In this task, the answer resides in a single sentence or paragraph within a long context, and no additional reasoning or computation is required to formulate a correct response, such as identifying a character’s name or a specific value mentioned in the text. It is worth noting that the location task differs from the “Needle in a Haystack” problem (Kamradt., 2023), which focuses on verbatim matching. In contrast, the location task allows for paraphrasing, as long as the underlying meaning is preserved. This task is crucial for assessing an LLM’s

basic comprehension and information retrieval capabilities within a long context.

Reasoning. The reasoning task in LaRA involves questions that require logical deduction, inference, or calculation based on the information provided in the long context. Instead of directly extracting answers from the text, these tasks demand a deeper understanding and processing of the information to derive the correct answer, such as inferring the relationship between two characters or calculating relevant data in financial statements. These tasks evaluate the ability of LC and RAG to handle complex questions, particularly in scenarios where the long context contains a significant amount of noise irrelevant to the question. The specific questions vary significantly depending on the type of context involved. Instead of explicitly defining sub-task types, we adopt different seed questions tailored to specific text types. These seed questions are used to generate similar QA pairs through in-context learning. For example, in financial statements, which contain a significant amount of statistical data, we focus on computational questions, and for novels, the questions involve reasoning about the plot or character traits.

Comparison. The comparison task in LaRA evaluates the ability of RAG and LC to synthesize information from multiple parts of a long context, comparing their content or numerical values to arrive at the final answer. Crucially, the comparison task also involves manually designing different seed questions tailored to various text types. This approach ensures that the generated questions are not only relevant but also reflect the nuances and complexities of the specific context. For instance, in academic papers, the questions may focus on comparing different explanations of the same phenomenon, while in novels, they may compare changes in a character’s traits or appearance over time. This task is essential for assessing an LLM’s ability to extract information from different parts.

Hallucination detection. Hallucination, a common issue in LLMs, occurs when the model generates inaccurate or irrelevant information (Huang et al., 2023). The hallucination detection task aims to test the model’s ability to decline answering questions that are not mentioned in the given context. Although the questions appear to be answerable using the context, the required information is not actually mentioned in the text. Consequently, such questions have a uniform answer: “XXX is not mentioned in the provided context.” The ability to refuse to answer is crucial in practical applications of RAG and LC, particularly in domains where accuracy and reliability are paramount, as users cannot always guarantee that their questions have answers within the provided context. For example, a user might pose a seemingly relevant question about a paper,

²<https://www.gutenberg.org/>

³<https://www.annualreports.com/>

⁴<https://arxiv.org/>

and if the model hallucinates and generates an incorrect response, it could be highly detrimental.

3.3. Data Annotation

The annotation process for different tasks follows a similar framework, starting with the manual creation of seed questions and answers. We then utilize GPT-4o to generate new QA pairs through in-context learning. A subset of newly generated QAs is sampled for manual validation to ensure correctness and practicality. If the pass rate does not meet a predefined threshold (we sample 40 generated QAs for each context type and task, and the threshold is set to 90%), the seed QAs and prompts are refined, followed by re-generation and re-validation. We provide the annotation prompt in Appendix E.

Annotating long texts presents a unique challenge due to the inherent difficulty of long context processing. One effective approach to improve generation quality is to convert annotations for long texts into annotations for shorter texts. To achieve this, we employ various strategies tailored to different context types and tasks. Specifically, for location and reasoning tasks, we split the long context into multiple segments, each approximately 10k tokens in length, and input them individually into GPT-4o to generate QAs. This approach serves multiple purposes: First, it reduces the cognitive load on the annotator (GPT-4o here) and improves the focus and accuracy of the generated QA pairs. Second, it ensures that the answers are evenly distributed across the entire context, as we observe that providing the full context to the LLM often results in answers being concentrated at the beginning and end of the context. Third, it allows us to examine the relationship between answer accuracy and answer location, enabling us to investigate whether the LLM suffers from the “lost in the middle” issue, where performance declines for information located in the middle sections of long documents (Liu et al., 2024a). For the comparison task, we split the context into smaller segments and then sample two segments to generate comparison questions similar to the seed questions. Meanwhile, our segmentation strategies are tailored to the specific context type to preserve the inherent structure and coherence of the documents. For research papers, we separate concatenated papers to maintain the integrity of each individual paper. For novels and financial statements, we directly split the text into multiple segments based on token count.

3.4. Evaluation

Metrics. Automated evaluation metrics, such as F1-score and ROUGE, can often produce lower scores, while LLM evaluations for QA tasks with definitive answers have demonstrated high precision (Chiang & Lee, 2023; Liu et al., 2023). Therefore, we provide GPT-4o with the query,

the ground-truth answer, and the prediction, enabling it to assess the correctness of the response (details and prompt are provided in Appendix C). Since LaRA consists solely of QA pairs with clearly defined answers and no open-ended questions, using LLM as a judge is highly effective in ensuring accuracy and consistency in the evaluation process.

Manual Verification. To ensure the quality and reliability of LaRA, we incorporate manual verification at two stages of the construction process. First, during the generation process, we employ a pipeline involving sampling, prompt refinement, and seed QA selection as manual adjustments. We find that the choice of seed questions has the most significant impact, possibly because LLMs perform much better in in-context learning compared to zero-shot question generation, demonstrating the importance of providing relevant examples for guiding the generation process (For details, see Appendix E). Second, we calculate the Cohen’s Kappa coefficient between the evaluation from LLM and human to quantitatively assess the agreement between the LLM and human evaluations, ensuring consistency and reliability in the judgment process. We provide the details in Appendix C.

4. Experiments

4.1. Experimental Settings

Baselines. To investigate the impact of various factors on RAG and LC performance, we evaluate a diverse set of 11 LLMs, encompassing both open-source and proprietary models. This includes seven open-source LLMs: Llama-3.2-3B-Instruct (Meta, 2024a), Llama-3.1-8B-Instruct (Meta, 2024b), Llama-3.3-70B-Instruct (Meta, 2024c), Llama-3.3-70B-Instruct-Q8 (Meta, 2024c) (utilizing FP8 quantization), Qwen-2.5-7B-Instruct (Yang et al., 2024), Qwen-2.5-72B-Instruct (Yang et al., 2024), and Mistral-Nemo-12B (AI, 2024). We also evaluate four advanced proprietary LLMs: GPT-4o (OpenAI, 2024), GPT-4o-mini (OpenAI, 2024), Claude-3.5-Sonnet (Anthropic, 2024), and Gemini-1.5-Pro (Reid et al., 2024).

Implementation of RAG. We employ a common configuration with a chunk size of 600 tokens, 5 chunks per document, and an overlap of 100 tokens between chunks. We utilize GTE-large-en-v1.5 (Zhang et al., 2024b; Li et al., 2023) for embedding extraction and reranking and employ a hybrid search strategy (Wang et al., 2024b) combining embedding similarity and BM25 (Robertson et al., 2009). Before conducting system experiments, we attempted to replace the embedding model and rerank model, such as with bge-m3 (Chen et al., 2024b), and also tried adding a summarization module (Xu et al., 2023). However, we found that these changes had minimal impact on the

Table 2. The accuracy of baselines evaluated by GPT-4o on LaRA (%). We evaluate performance for context lengths of 32k and 128k tokens separately, with the gray background representing RAG and the white background representing LC. The highest-performing open-source and proprietary LLMs for each task are highlighted in bold. “Avg GAP” refers to the difference between the average accuracy of LC and RAG across all models for a specific task (calculated as LC minus RAG). Blue text indicates that LC performs better, while red text indicates that RAG is better.

Model	Location		Reasoning		Comparison		Hallucination		Overall	
Open-source LLMs (32k)										
Llama-3.1-8B-Instruct	77.60	73.85	51.65	46.59	58.29	38.06	78.80	83.73	66.58	60.56
Llama-3.2-3B-Instruct	69.11	71.08	36.60	34.88	31.10	25.02	65.13	79.71	50.48	52.67
Llama-3.3-70B-Instruct	77.50	79.83	66.43	57.23	65.33	53.13	73.59	86.62	70.71	69.20
Llama-3.3-70B-Instruct-Q8	78.24	77.86	61.12	58.55	64.83	47.90	74.09	83.95	69.57	67.06
Qwen-2.5-7B-Instruct	78.10	73.70	51.30	48.63	62.35	43.79	76.42	84.36	67.04	62.62
Qwen-2.5-72B-Instruct	83.29	81.63	74.06	65.24	68.89	49.24	85.08	83.76	77.83	69.97
Mistral-Nemo-12B	54.45	72.38	29.81	45.74	29.02	43.63	35.27	71.70	37.14	58.36
Proprietary LLMs (32k)										
GPT-4o	86.33	82.55	75.48	68.51	79.66	46.56	72.30	78.39	78.44	69.00
GPT-4o-mini	79.92	77.69	77.65	62.23	73.91	54.48	52.82	67.17	71.08	65.39
Claude-3.5-sonnet	85.82	83.10	70.42	66.81	66.28	51.38	86.12	90.99	77.16	73.07
Gemini-1.5-pro	78.94	74.71	64.68	53.50	77.34	56.43	84.58	88.03	76.39	68.17
Avg GAP	0.08		4.66		15.22		-10.38		2.40	
Open-source LLMs (128k)										
Llama-3.1-8B-Instruct	72.64	72.65	48.48	47.73	40.87	22.27	58.40	78.36	55.10	55.25
Llama-3.2-3B-Instruct	60.96	68.96	33.99	38.86	24.54	20.20	52.06	69.45	42.89	49.37
Llama-3.3-70B-Instruct	74.54	78.11	52.98	59.05	43.09	26.69	44.00	77.67	53.65	60.38
Llama-3.3-70B-Instruct-Q8	72.44	77.99	53.97	58.32	45.32	29.61	38.18	76.87	52.48	60.70
Qwen-2.5-7B-Instruct	68.94	74.08	44.69	50.93	39.51	29.17	42.52	71.02	48.91	56.30
Qwen-2.5-72B-Instruct	76.10	78.92	65.25	64.37	54.62	36.10	64.45	71.32	65.11	62.68
Mistral-Nemo-12B	23.44	72.29	14.87	50.53	7.77	32.59	14.77	57.92	15.21	53.33
Proprietary LLMs (128k)										
GPT-4o	87.70	82.22	79.35	70.26	64.74	40.84	56.34	67.33	72.03	65.16
GPT-4o-mini	79.86	80.43	67.80	65.15	65.31	41.29	32.08	59.31	61.26	61.55
Claude-3.5-sonnet	85.44	80.94	73.79	64.81	58.35	33.18	77.64	84.73	73.81	65.92
Gemini-1.5-pro	82.15	75.41	67.97	48.47	61.67	36.54	66.16	78.60	69.49	59.75
Avg GAP	-5.25		-1.39		14.30		-22.36		-3.68	

performance and could potentially introduce additional computational overhead.

4.2. Main Results and Analysis

Overall performance. Table 2 presents the evaluation results, comparing the performance of LC and RAG across various LLMs. The table uses a white background for LC results and a gray background for RAG results, with the highest-performing open-source and proprietary LLMs for each task highlighted in bold. Our analysis reveals a complex relationship between model architecture, context length, and performance. For open-source models at a 32k context length, LC generally outperforms RAG, with the exception of Llama-3.2-3B-Instruct and Mistral-Nemo-12B. However, this trend reverses at a 128k context length,

where RAG demonstrates superior performance across most models. In contrast, proprietary models consistently favor LC at both context lengths, likely due to their larger parameter sizes and enhanced ability to process long-context inputs. The inherent self-attention mechanism in these models appears more effective at handling extended contexts compared to the sparse attention employed in RAG. Notably, at a 128k context length, the top three performing models (GPT-4o, Gemini-1.5-Pro, and Claude-3.5-Sonnet) all utilize LC, while the bottom three (Llama-3.2-3B-Instruct, Qwen-2.5-7B-Instruct, and Mistral-Nemo-12B) are also LC-based. This observation underscores the absence of a universal “winner” between RAG and LC, as performance is highly dependent on the specific LLM and context length.

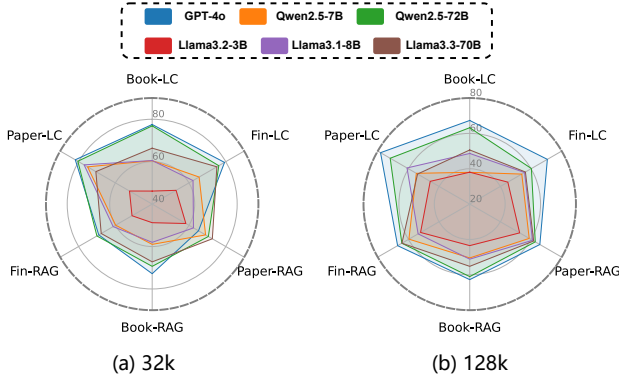


Figure 1. The average accuracy across different context types. The left figure (a) represents a context length of 32k, while the right figure (b) represents a context length of 128k.

Scaling law holds in LC. Our experimental results confirm the established scaling law in LC (Kaplan et al., 2020): larger models consistently outperform smaller counterparts. For example, GPT-4o and Qwen-2.5-72B-Instruct show significant performance gains ranging from 7.35% to 16.2% over their smaller versions, GPT-4o-mini and Qwen-2.5-7B-Instruct, respectively. This advantage is further amplified at a 128k context length. While all models experience a performance decline with longer contexts, the drop is more pronounced for smaller models, highlighting their limitations in processing extensive textual input. This observation challenges the purported ability of smaller models to handle extremely long contexts effectively.

RAG empowers models to handle extremely long context. At a 128k context length, RAG consistently outperforms LC across nearly all open-source models. For example, Llama-3.1-8B-Instruct and Qwen-2.5-7B-Instruct demonstrate improvements of 0.15% and 7.39%, respectively, when using RAG instead of LC. All models exhibit a performance decline at a 128k context length compared to 32k. However, LC experiences a more significant drop than RAG, indicating that as context length approaches its limit, RAG is less affected by the increase in context length. Furthermore, RAG enables models with weaker long-context capabilities, such as Mistral-Nemo-13B and Llama-3.2-3B-Instruct, to achieve performance comparable to other models. These findings highlight that while larger models excel at long-context processing, RAG offers an effective alternative for smaller or weaker models, ensuring competitive performance even with extended context lengths.

4.3. Task Analysis

Table 2 presents a detailed breakdown of LC and RAG performance across four distinct tasks. To provide a clear overview of the relative strengths of each task, we also

calculate the average performance gap (avg gap) for each task, representing the mean difference in accuracy between LC and RAG across all evaluated models.

Location. The location task proves to be the easiest among the four, with both RAG and LC achieving high accuracy. The average performance gap between RAG and LC is 0.08% at a 32k context length and -5.25% at 128k, indicating a slight advantage for LC with shorter contexts and a more pronounced advantage for RAG with longer contexts. For open-source models at 32k, the performance difference between RAG and LC is minimal, while at 128k, RAG demonstrates superior performance. This suggests that when models struggle to process long texts, retrieval acts as a valuable tool for location-based questions. Conversely, for proprietary models, LC consistently outperforms RAG, indicating that with sufficient model capacity, LLMs can outperform RAG in handling such simple tasks on their own.

Reasoning. The performance trends for reasoning tasks basically mirror those observed in the location tasks, particularly for smaller models. With a 32k context, RAG exhibits slightly lower accuracy compared to LC, while at 128k, this trend reverses. However, for larger models like GPT-4o and Claude-3.5-Sonnet, the advantage of LC becomes more pronounced. At a 128k context length, GPT-4o and Claude-3.5-Sonnet outperform RAG by 9.09% and 8.98% in accuracy, respectively. We speculate that while reasoning tasks often rely on specific text segments for answers, other parts of the document may contain supplementary information that aids in inference. Models with stronger long-context capabilities are better equipped to leverage this global knowledge, leading to improved performance in reasoning tasks.

Comparison. Comparative tasks pose the greatest challenge for RAG, showing the largest performance gap compared to LC. The average gap reaches 15.22% at 32k and 14.30% at 128k. A deeper analysis of problematic cases reveals two primary reasons for the struggle of RAG. First, some comparative questions emphasize one aspect of the comparison while providing limited information about the other, making it difficult for RAG to retrieve both necessary chunks for a correct answer. Second, certain queries describe the comparison in abstract terms rather than concrete details, hindering the effectiveness of similarity-based retrieval. This abstraction makes it challenging for RAG to locate all the relevant chunks for comparison. In contrast to location tasks, which involve pinpointing a single piece of information, comparative tasks require the accurate retrieval and comparison of multiple distinct chunks, significantly increasing the complexity for RAG.

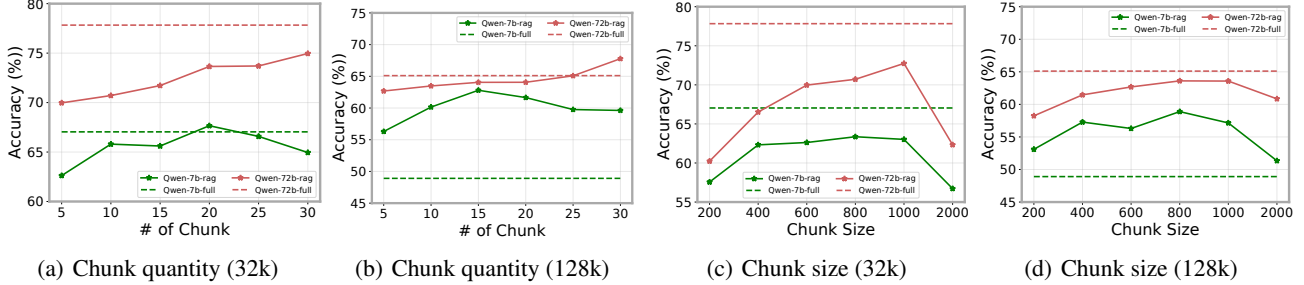


Figure 2. The accuracy of Qwen-2.5-72B-Instruct and Qwen-2.5-7B-Instruct with different chunk quantity and size on LaRA.

Hallucination detection. This is the only task where RAG demonstrates a clear advantage in both small and large models. LC tends to generate more hallucinated or incorrect answers, likely due to the increased noise introduced by feeding the entire text to the model. This makes the model more susceptible to errors and distractions, leading to fabricated responses. Interestingly, larger models do not exhibit an evident advantage in this task. While GPT-4o achieves top performance on other tasks, it only attains an accuracy of 56.34% on hallucination detection. This suggests that even models capable of handling long contexts can be overwhelmed by the sheer volume of information and generate flawed conclusions. In contrast, RAG’s selective retrieval of relevant information helps mitigate this issue, enabling both small and large models to better identify situations where they lack sufficient knowledge to provide an accurate answer.

4.4. Context Type Analysis

We present the influence of different context types on the performance of RAG and LC in Figure 1. Novel-related questions pose the greatest challenge, while paper-related questions are the easiest. This disparity likely stems from the repetitive sentence structures common in novels, which can hinder precise information localization. Conversely, academic papers typically exhibit a stronger logical flow and higher information density, facilitating easier distinction between questions and answers. At a 32k context length, LC outperforms RAG for nearly all models. However, at 128k, weaker models demonstrate better performance with RAG. Interestingly, the performance gap between RAG and LC is smaller for novel-related tasks compared to paper-related or financial statements tasks, regardless of context length. This suggests that for less structured contexts, RAG presents a viable alternative for reducing computational cost. On the other hand, for highly structured texts like academic papers and financial statements, LC demonstrates a clear advantage.

4.5. Impact of Chunk Quantity and Size

We explore the impact of the length of retrieved information length on RAG performance through two aspects: the

number of chunks and the size of each chunk. We conduct experiments on Qwen-2.5-72B-Instruct and Qwen-2.5-7B-Instruct to observe the impact of chunk size and quantity on both large and small models. As shown in Figure 2, for the 72B model, performance improves consistently as the number of retrieved chunks increases, benefiting from its stronger long-context processing capability. In contrast, the 7B model exhibits a peak performance at an intermediate chunk quantity, after which excessive retrieval introduces noise that outweighs the information gain. Regarding chunk size, both excessively large and small chunks lead to performance degradation. Within a reasonable range, increasing chunk size provides some improvement, though its effect is less significant than increasing the number of chunks.

4.6. Lost in The Middle

By controlling the location of the information required to answer the questions, we find that LC LLMs exhibit a decrease in accuracy when the answer is closer to the center of the context, indicating a susceptibility to the “lost in the middle” phenomenon. In contrast, we do not observe a clear correlation between RAG performance and the position of the answer, suggesting that RAG models are more robust to this issue. These findings highlight a potential advantage of RAG models in handling long contexts, particularly for tasks that require accessing information from various parts of the document. For detailed results, see Appendix D.

5. Conclusion

This study addressed the critical question of whether RAG or LC is more effective for incorporating external knowledge into LLMs. Through the development and evaluation of LaRA, a novel benchmark, we demonstrated that the optimal choice depends on a complex interplay of factors, including model size, context length, and task type. Our findings challenge previous inconclusive comparisons and offer actionable guidelines for practitioners. LaRA serves as a valuable resource for evaluating and comparing RAG and LC models, facilitating further research in this rapidly evolving field.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

- AI, M. Mistral large, 2024. URL <https://mistral.ai/news/mistral-nemo/>.
- Anthropic. Claude 3.5 sonnet, 2024. URL <https://www.anthropic.com/news/claude-3-5-sonnet/>.
- 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. Longbench: A bilingual, multitask benchmark for long context understanding. In Ku, L., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, pp. 3119–3137. Association for Computational Linguistics, 2024a. doi: 10.18653/V1/2024.ACL-LONG.172. URL <https://doi.org/10.18653/v1/2024.acl-long.172>.
- Bai, Y., Tu, S., Zhang, J., Peng, H., Wang, X., Lv, X., Cao, S., Xu, J., Hou, L., Dong, Y., et al. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *arXiv preprint arXiv:2412.15204*, 2024b.
- Chen, J., Lin, H., Han, X., and Sun, L. Benchmarking large language models in retrieval-augmented generation. In Wooldridge, M. J., Dy, J. G., and Natarajan, S. (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pp. 17754–17762. AAAI Press, 2024a. doi: 10.1609/AAAI.V38I16.29728. URL <https://doi.org/10.1609/aaai.v38i16.29728>.
- Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., and Liu, Z. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024b.
- Chiang, D. C. and Lee, H. Can large language models be an alternative to human evaluations? In Rogers, A., Boyd-Graber, J. L., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2023, Toronto, Canada, July 9-14, 2023, pp. 15607–15631. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.870. URL <https://doi.org/10.18653/v1/2023.acl-long.870>.
- Dasigi, P., Lo, K., Beltagy, I., Cohan, A., Smith, N. A., and Gardner, M. A dataset of information-seeking questions and answers anchored in research papers. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tür, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y. (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pp. 4599–4610. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.NAACL-MAIN.365. URL <https://doi.org/10.18653/v1/2021.naacl-main.365>.
- Dong, Z., Tang, T., Li, J., Zhao, W. X., and Wen, J. BAMBOO: A comprehensive benchmark for evaluating long text modeling capacities of large language models. In Calzolari, N., Kan, M., Hoste, V., Lenci, A., Sakti, S., and Xue, N. (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pp. 2086–2099. ELRA and ICCL, 2024. URL <https://aclanthology.org/2024.lrec-main.188>.
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M., and Wang, H. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997, 2023. doi: 10.48550/ARXIV.2312.10997. URL <https://doi.org/10.48550/arXiv.2312.10997>.
- Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3929–3938. PMLR, 2020. URL <http://proceedings.mlr.press/v119/guu20a.html>.
- Hsieh, C., Sun, S., Krizan, S., Acharya, S., Rekish, D., Jia, F., Zhang, Y., and Ginsburg, B. RULER: what’s the real context size of your long-context language models? *CoRR*, abs/2404.06654, 2024. doi: 10.48550/ARXIV.2404.06654. URL <https://doi.org/10.48550/arXiv.2404.06654>.
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu,

- T. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *CoRR*, abs/2311.05232, 2023. doi: 10.48550/ARXIV.2311.05232. URL <https://doi.org/10.48550/arXiv.2311.05232>.
- Kamradt, G. Needle in a haystack - pressure testing llms., 2023. URL https://github.com/gkamradt/LLMTest_NeedleInAHaystack.
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020. URL <https://arxiv.org/abs/2001.08361>.
- Kociský, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K. M., Melis, G., and Grefenstette, E. The narrativeqa reading comprehension challenge. *Trans. Assoc. Comput. Linguistics*, 6:317–328, 2018. doi: 10.1162/TACL\A\00023. URL https://doi.org/10.1162/tacl_a_00023.
- Lee, J., Chen, A., Dai, Z., Dua, D., Sachan, D. S., Boratko, M., Luan, Y., Arnold, S. M. R., Perot, V., Dalmia, S., Hu, H., Lin, X., Pasupat, P., Amini, A., Cole, J. R., Riedel, S., Naim, I., Chang, M., and Guu, K. Can long-context language models subsume retrieval, rag, sql, and more? *CoRR*, abs/2406.13121, 2024. doi: 10.48550/ARXIV.2406.13121. URL <https://doi.org/10.48550/arXiv.2406.13121>.
- Lewis, P. S. H., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>.
- Li, J., Sun, A., Han, J., and Li, C. A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.*, 34(1):50–70, 2022. doi: 10.1109/TKDE.2020.2981314. URL <https://doi.org/10.1109/TKDE.2020.2981314>.
- Li, J., Wang, M., Zheng, Z., and Zhang, M. Loogle: Can long-context language models understand long contexts? In Ku, L., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, pp. 16304–16333. Association for Computational Linguistics, 2024a. doi: 10.18653/V1/2024.ACL-LONG.859. URL <https://doi.org/10.18653/v1/2024.acl-long.859>.
- Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., and Zhang, M. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023.
- Li, Z., Li, C., Zhang, M., Mei, Q., and Bendersky, M. Retrieval augmented generation or long-context llms? A comprehensive study and hybrid approach. In Dernoncourt, F., Preotiuc-Pietro, D., and Shimorina, A. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: EMNLP 2024 - Industry Track, Miami, Florida, USA, November 12-16, 2024*, pp. 881–893. Association for Computational Linguistics, 2024b. URL <https://aclanthology.org/2024.emnlp-industry.66>.
- Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguistics*, 12:157–173, 2024a. doi: 10.1162/TACL\A\00638. URL https://doi.org/10.1162/tacl_a_00638.
- Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. G-eval: NLG evaluation using gpt-4 with better human alignment. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 2511–2522. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.153. URL <https://doi.org/10.18653/v1/2023.emnlp-main.153>.
- Liu, Y., Yang, T., Huang, S., Zhang, Z., Huang, H., Wei, F., Deng, W., Sun, F., and Zhang, Q. Calibrating llm-based evaluator. In Calzolari, N., Kan, M., Hoste, V., Lenci, A., Sakti, S., and Xue, N. (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pp. 2638–2656. ELRA and ICCL, 2024b. URL <https://aclanthology.org/2024.lrec-main.237>.
- Long, L., Wang, R., Xiao, R., Zhao, J., Ding, X., Chen, G., and Wang, H. On llms-driven synthetic data generation, curation, and evaluation: A survey. In Ku, L., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok,*

- Thailand and virtual meeting, August 11-16, 2024*, pp. 11065–11082. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.658. URL <https://doi.org/10.18653/v1/2024.findings-acl.658>.
- Lyu, Y., Li, Z., Niu, S., Xiong, F., Tang, B., Wang, W., Wu, H., Liu, H., Xu, T., and Chen, E. CRU-RAG: A comprehensive chinese benchmark for retrieval-augmented generation of large language models. *CoRR*, abs/2401.17043, 2024. doi: 10.48550/ARXIV.2401.17043. URL <https://doi.org/10.48550/arXiv.2401.17043>.
- Meta. Llama 3.2, 2024a. URL <https://ai.meta.com>.
- Meta. Llama 3.1, 2024b. URL <https://ai.meta.com/blog/meta-llama-3-1/>.
- Meta. Llama 3.3, 2024c. URL https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/.
- Novikova, J., Dusek, O., Curry, A. C., and Rieser, V. Why we need new evaluation metrics for NLG. In Palmer, M., Hwa, R., and Riedel, S. (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pp. 2241–2252. Association for Computational Linguistics, 2017. doi: 10.18653/V1/D17-1238. URL <https://doi.org/10.18653/v1/d17-1238>.
- OpenAI. Gpt-4o, 2024. URL <https://openai.com/index/hello-gpt-4o/>.
- Pang, R. Y., Parrish, A., Joshi, N., Nangia, N., Phang, J., Chen, A., Padmakumar, V., Ma, J., Thompson, J., He, H., and Bowman, S. R. Quality: Question answering with long input texts, yes! In Carpuat, M., de Marneffe, M., and Ruíz, I. V. M. (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pp. 5336–5358. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.NAACL-MAIN.391. URL <https://doi.org/10.18653/v1/2022.naacl-main.391>.
- Qi, Z., Xu, R., Guo, Z., Wang, C., Zhang, H., and Xu, W. Long2rag: Evaluating long-context & long-form retrieval-augmented generation with key point recall. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4852–4872, 2024.
- Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T. P., Alayrac, J., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., Antonoglou, I., Anil, R., Borgeaud, S., Dai, A. M., Millican, K., Dyer, E., Glaese, M., Sottiaux, T., Lee, B., Viola, F., Reynolds, M., Xu, Y., Molloy, J., Chen, J., Isard, M., Barham, P., Hennigan, T., McIlroy, R., Johnson, M., Schalkwyk, J., Collins, E., Rutherford, E., Moreira, E., Ayoub, K., Goel, M., Meyer, C., Thornton, G., Yang, Z., Michalewski, H., Abbas, Z., Schucher, N., Anand, A., Ives, R., Keeling, J., Lenc, K., Haykal, S., Shakeri, S., Shyam, P., Chowdhery, A., Ring, R., Spencer, S., Sezener, E., and et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530, 2024. doi: 10.48550/ARXIV.2403.05530. URL <https://doi.org/10.48550/arXiv.2403.05530>.
- Robertson, S., Zaragoza, H., et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.
- Saad-Falcon, J., Khattab, O., Potts, C., and Zaharia, M. ARES: an automated evaluation framework for retrieval-augmented generation systems. In Duh, K., Gómez-Adorno, 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), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pp. 338–354. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.20. URL <https://doi.org/10.18653/v1/2024.naacl-long.20>.
- Shaham, U., Ivgi, M., Efrat, A., Berant, J., and Levy, O. Zeroscrolls: A zero-shot benchmark for long text understanding. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pp. 7977–7989. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-EMNLP.536. URL <https://doi.org/10.18653/v1/2023.findings-emnlp.536>.
- Stolfo, A. Groundedness in retrieval-augmented long-form generation: An empirical study. In Duh, K., Gómez-Adorno, H., and Bethard, S. (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pp. 1537–1552. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-NAACL.100. URL <https://doi.org/10.18653/v1/2024.findings-naacl.100>.

- Wang, M., Chen, L., Cheng, F., Liao, S., Zhang, X., Wu, B., Yu, H., Xu, N., Zhang, L., Luo, R., Li, Y., Yang, M., Huang, F., and Li, Y. Leave no document behind: Benchmarking long-context llms with extended multi-doc QA. In Al-Onaizan, Y., Bansal, M., and Chen, Y. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 5627–5646. Association for Computational Linguistics, 2024a. URL <https://aclanthology.org/2024.emnlp-main.322>.
- Wang, X., Jiang, Y., Bach, N., Wang, T., Huang, Z., Huang, F., and Tu, K. Improving named entity recognition by external context retrieving and cooperative learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1800–1812, Online, August 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.acl-long.142>.
- Wang, X., Shen, Y., Cai, J., Wang, T., Wang, X., Xie, P., Huang, F., Lu, W., Zhuang, Y., Tu, K., Lu, W., and Jiang, Y. DAMO-NLP at SemEval-2022 task 11: A knowledge-based system for multilingual named entity recognition. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pp. 1457–1468, Seattle, United States, July 2022. Association for Computational Linguistics. URL <https://aclanthology.org/2022.semeval-1.200>.
- Wang, X., Wang, Z., Gao, X., Zhang, F., Wu, Y., Xu, Z., Shi, T., Wang, Z., Li, S., Qian, Q., et al. Searching for best practices in retrieval-augmented generation. *arXiv preprint arXiv:2407.01219*, 2024b.
- Xu, F., Shi, W., and Choi, E. Recomp: Improving retrieval-augmented lms with compression and selective augmentation. *arXiv preprint arXiv:2310.04408*, 2023.
- Xu, P., Ping, W., Wu, X., McAfee, L., Zhu, C., Liu, Z., Subramanian, S., Bakhturina, E., Shoenybi, M., and Catanzaro, B. Retrieval meets long context large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=xw5nxFWMlo>.
- Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- Yu, T., Xu, A., and Akkiraju, R. In defense of RAG in the era of long-context language models. *CoRR*, abs/2409.01666, 2024. doi: 10.48550/ARXIV.2409.01666. URL <https://doi.org/10.48550/arXiv.2409.01666>.
- 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. inftybench: Extending long context evaluation beyond 100k tokens. In Ku, L., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 15262–15277. Association for Computational Linguistics, 2024a. doi: 10.18653/V1/2024.ACL-LONG.814. URL <https://doi.org/10.18653/v1/2024.acl-long.814>.
- Zhang, X., Zhang, Y., Long, D., Xie, W., Dai, Z., Tang, J., Lin, H., Yang, B., Xie, P., Huang, F., et al. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. *arXiv preprint arXiv:2407.19669*, 2024b.
- Zhang, Z., Fang, M., and Chen, L. Retrievalqa: Assessing adaptive retrieval-augmented generation for short-form open-domain question answering. In Ku, L., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pp. 6963–6975. Association for Computational Linguistics, 2024c. doi: 10.18653/V1/2024.FINDINGS-ACL.415. URL <https://doi.org/10.18653/v1/2024.findings-acl.415>.

A. Statistics of LaRA

LaRA consists of approximately 2300 test cases, encompassing three context types and four task categories. To ensure that the token count of the context is as close as possible to 32k and 128k without exceeding these limits, the primary token ranges for these two lengths are 20-30k and 80-120k, respectively. The average token counts and numbers of tasks are provided in Table 3.

Table 3. Statistics of LaRA.

Context	Location	Reasoning	Comparison	Hallucination
32k length				
Novel	25673	25908	25681	25433
Financial	27548	27531	27546	27527
Paper	28078	28088	27708	28081
# of Cases	276	230	151	230
128k length				
Novel	96452	96226	95903	96182
Financial	92684	92831	92830	92812
Paper	93911	93818	94731	93890
# of Cases	489	374	198	378

B. Named Entity Recognition and Replacement for Novels

We initially tried using some traditional Named Entity Recognition (NER) methods (Wang et al., 2021; 2022), but found that their performance was poor. First, traditional sequence labeling models struggle with long-text processing due to fixed-length context windows, failing to capture cross-paragraph entity associations. Second, performance degrades significantly on out-of-distribution data, particularly when handling domain-specific or stylistically unique texts (Li et al., 2022). These constraints prove especially problematic for literary analysis, where novels exhibit both long-range narrative dependencies and rich variations in entity references (e.g., honorifics, epithets, and contextual substitutions).

The emergence of LLM presents new opportunities to overcome these limitations through their superior contextual understanding and robust generalization capabilities. Hence, we leverage GPT-4o to perform entity extraction and replacement in full-length novels through a three-stage pipeline:

1. We partition input texts into coherent segments averaging 500 tokens, preserving complete sentence/paragraph boundaries. This chunk size optimizes the balance between LLMs’ context window constraints and narrative continuity requirements.
2. Each text chunk undergoes parallel entity recognition through GPT-4o processing. Our extraction protocol captures both original names and contextual variants (e.g., “The Dark Lord” → “Voldemort” in Harry Potter). After feeding each chunk for extraction, we merge all the entities and remove duplicates to obtain a final list of entities while preserving legitimate aliases.
3. For alias disambiguation, we prompt GPT-4o with the entity list and the novel’s title to determine if multiple names indicate the same character, and if they do, we replace them with the same fake name.

Figure 3 and Figure 4 detail our carefully engineered prompts for entity extraction and substitution respectively. Note that the novels we select are classical literary works, which, along with extensive related discussions, are included in the LLMs’ training data. As a result, we find that simply providing the name of the novel is sufficient to accurately map multiple aliases to the same character.

Example prompt used for *Named Entity Extraction*

Task: Your task is to extract named entities(only person) from the given paragraph. Response with a list of entities, and strings in the list should be enclosed in double quote. Below is an example, and you need to use a style consistent with the example.

Example:

<paragraph>Filby became pensive. “Clearly,” the Time Traveller proceeded, “any real body must have extension in four directions: it must have Length, Breadth, Thickness, and—Duration. But through a natural infirmity of the flesh, which I will explain to you in a moment, we incline to overlook this fact. There are really four dimensions, three which we call the three planes of Space, and a fourth, Time. There is, however, a tendency to draw an unreal distinction between the former three dimensions and the latter, because it happens that our consciousness moves intermittently in one direction along the latter from the beginning to the end of our lives.”

“That,” said a very young man, making spasmodic efforts to relight his cigar over the lamp; “that . . . very clear indeed.”

“Now, it is very remarkable that this is so extensively overlooked,” continued the Time Traveller, with a slight accession of cheerfulness. “Really this is what is meant by the Fourth Dimension, though some people who talk about the Fourth Dimension do not know they mean it. It is only another way of looking at Time. There is no difference between Time and any of the three dimensions of Space except that our consciousness moves along it. But some foolish people have got hold of the wrong side of that idea. You have all heard what they have to say about this Fourth Dimension?”

“I have not,” said the Provincial Mayor.</paragraph>

Entities: [“Filby”, “Time Traveller”, “Provincial Mayor”]

<paragraph>{**chunk**}</paragraph>

Entities:

Figure 3. The prompt for extracting named entities from novel chunks. The chunk that need to be extracted is highlighted in red text.

C. LLM as A Judge

Given the unreliability of rule-based evaluations and the high costs associated with human evaluation, the use of LLM for assessment has gained increasing popularity (Liu et al., 2024b; Wang et al., 2024a). In LaRA, we prompt GPT-4o to determine whether a model correctly answer a question, using the query, the ground-truth answer, and the model’s prediction as inputs. The specific prompt is shown in Figure 5.

To verify the consistency between GPT-4o evaluation and human evaluation, we compute the Cohen’s Kappa coefficient, which is defined as:

$$\kappa = \frac{P_o - P_e}{1 - P_e}, \quad (1)$$

where P_o represents the proportion of agreement between the two evaluators on the positive and negative classes, and P_e denotes the probability of agreement between evaluators under the assumption of independent and random classification.:

$$P_o = \frac{TP + TN}{TP + FP + TN + FN}. \quad (2)$$

$$P_e = \frac{(TP + FN)(TP + FP) + (TN + FN)(TN + FP)}{(TP + FP + TN + FN)^2}, \quad (3)$$

where TP and TN refer to cases where both the human evaluator and the model agree that the answer is correct, while FP and FN refer to cases where only one of them judges the answer as correct.

Example prompt used for *Name Replacement*

Task: I have extracted the names of all the characters in the novel {**novel**}, as shown in the list below. Your task is to replace the names in a novel with fictitious ones.

Firstly, You need to determine based on the novel’s content whether any of these names refer to the same character. If they refer to the same person, they should be assigned the same fictitious name. You should respond with a Python dict only, with the keys being the names from the novel and the values being the new fictitious names. Strings in the dict should be enclosed in double quote

Name list: {**name_list**}

Figure 4. The prompt for replacing the names with fictitious ones.

Example prompt used for *Evaluation*

Task: You are a discriminator that judges whether the predictions to questions are correct.

I will provide you with a question and its Ground-truth answer, as well as an answer from an AI assistant. You need to judge whether the AI assistant’s answer is correct based on the Ground-truth answer. If it is correct, you should only output True; if it is incorrect, only output False.

[Query] {**query**}

[Ground-truth Answer] {**label**}

[AI Assistant’s Answer] {**pred**}

Your judgment:

Figure 5. The prompt for evaluation.

A Cohen’s Kappa coefficient greater than zero indicates consistency in the evaluation, with values approaching 1 indicating stronger agreement. We sample 100 predictions from each task type at a 128k context length from GPT-4o (LC) and Qwen-2.5-7B (LC) for manual evaluation and then compute the Cohen’s Kappa coefficient. The results are provided in Table 4. The values for both models are close to 1, demonstrating that evaluations from GPT-4o are highly consistent with human evaluation, whether applied to large or small models.

Table 4. Cohen’s Kappa coefficient for GPT-4o (LC) and Qwen-2.5-7B (LC).

Model	Location	Reasoning	Comparison	Hallucination
Qwen-2.5-7B	0.9737	0.9184	0.8955	0.9800
GPT-4o	0.9509	0.9254	0.9776	0.9604

D. Lost in The Middle

To assess the presence of “lost in the middle” in our LaRA benchmark, we ensure a uniform distribution of answers across different positions within the context during the annotation of location and reasoning tasks. Specifically, for location tasks

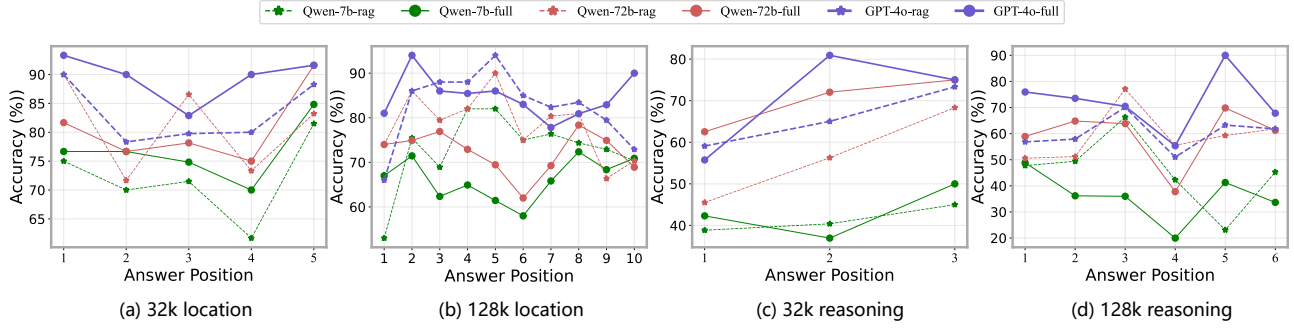


Figure 6. The accuracy of the location and reasoning tasks when the answer appears at different positions within the context. 32k-length and 128k-length contexts are split into 5 and 10 segments, respectively. The contexts here are novels and financial statements, as we split papers in another method mentioned in Section 3.3.

with 32k and 128k context lengths, we divide the context into 5 and 10 segments, respectively, and for reasoning tasks, we use 3 and 6 segments for 32k and 128k contexts, respectively.

As illustrated in Figure 6, our experiments reveal that LC models indeed suffer from the “lost in the middle” phenomenon. This issue is particularly pronounced in weaker models like Qwen-2.5-7B. In contrast, RAG demonstrates consistent accuracy regardless of answer position, showcasing its robustness against this phenomenon. This highlights a key advantage of RAG: its ability to effectively retrieve and utilize relevant information regardless of its location within the context.

E. Annotation Details

Recent advancements in LLMs have demonstrated their ability to generate high-quality text comparable to human output, positioning synthetic data generated by LLMs as a viable alternative or complement to human-generated data (Long et al., 2024). Benefiting from this, we use GPT-4o to generate highly diverse question-answer pairs across multiple scenarios and tasks.

Specifically, for each context type and task category, we design different prompts and seed questions based on distinct guiding principles. To better understand how we generate QA pairs in LaRA, we present the prompts used for generating financial statement QAs as an example. The prompts for the Location, Reasoning, Comparison, and Hallucination tasks are shown in Figures 7, 8, 9, and 10, respectively.

Each task’s annotations share some common principles, such as ensuring that questions are unanswerable without the provided context and keeping responses concise to avoid lengthy elaborations. Additionally, there are specific requirements to more precisely define the scope of the questions. For example, in the location task, we require the answer to be directly extracted from the context and prohibit any interpretive phrasing. In contrast, for the reasoning task, the answer must be impossible to obtain through simple lookup. Finally, seed example questions are provided for in-context learning, both as a reference and to ensure clear output formatting for easier subsequent processing. During fine-tuning prompts and seed questions, we find that once the problem types are precisely defined, the seed examples have a significant impact on the quality of the generated QA pairs. The higher the diversity of seed examples, the less uniform the generated questions become. Therefore, we strive to ensure that the seed questions cover various aspects and reflect the practical questions that users might realistically ask in real-world scenarios. For each task, we provide 3-5 seed examples.

F. Task Examples

For each task type, we provide three test examples, one for each type of context, to specifically demonstrate queries in LaRA that are close to real-world scenarios (Figure 11, 12, 13, and 14). Although we do not explicitly define lower-level sub-tasks within the four main task categories, it is evident from the examples that the emphasis of the same task varies across different context types. For example, financial statements are well-suited for mathematical calculations, while research papers are ideal for comparing different viewpoints and data.

We believe the most essential principle that all questions must adhere to is practicality—that is, the questions should reflect

Example prompt used for generating *Location* QAs in financial statements

Task: You are a financial question designer. Based on the provided financial statement, create a factual question and its corresponding answer that strictly satisfies these criteria:

Design Requirements:

1. Direct Extraction:
 - The answer must be explicitly stated in a single contiguous segment of the document
 - Require pinpoint localization (e.g., specific value, exact date, named section)
2. Answerability Constraints:
 - Unanswerable without the provided context
 - Must reference concrete elements (numerical values, named metrics, verbatim terms)
3. Response Specifications:
 - Maximum answer length: 20 words
 - Prohibit any interpretive phrasing

Example QAs:

{Q: “What was the amount of impairment changes recorded to goodwill during the three months ended December 31, 2023?”, A: “\$62.8 million.”}

{Q: “How many reportable business segments does 2U, Inc. have?”, A: “2U, Inc. has two reportable business segments.”}

...

Financial Statement: {Context}

Figure 7. The prompt for generating QA pairs.

those that a person could realistically ask based on the given context. While questions designed solely to challenge LLMs have academic value, they do not address the considerations necessary for designing a real-world RAG or LC planning system. *Our work aims to explore how to make more optimal choices under a realistic query distribution.*

Example prompt used for generating *Reasoning* QAs in financial statements

Task: You are a financial analyst. Design a question requiring mathematical/logical derivation from the financial statement, adhering to:

Design Requirements:

1. Require multi-step processing of:
 - Numerical calculations (e.g., ratios, growth rates)
 - Temporal comparisons
 - Reasonable inferences
2. Answerability Constraints:
 - Unanswerable without the provided context
 - Impossible to answer through simple lookup
3. Response Specifications:
 - Maximum answer length: avoid too long
 - Forbid speculative or probabilistic responses

Example QAs:

{Q: “Calculate the percentage increase in total assets from December 31, 2023, to March 31, 2024, and explain what this increase indicates about the company’s financial position during this period.”, A: “Total assets increased from \$22,008,739 on December 31, 2023, to \$36,852,097 on March 31, 2024, which is an increase of \$14,843,358. The percentage increase is $(14,843,358 / 22,008,739) * 100$, which equals approximately 67.5%. This significant increase in total assets indicates improved financial strength, potentially from additional capital inflows or asset acquisitions during the period.”}

...

Financial Statement: {Context}

Figure 8. The prompt for generating QA pairs.

Example prompt used for generating *Comparison* QAs in financial statements

Task: You are a financial analyst. Create a comparative question requiring synthesis of two distinct sections from the financial statement, following these principles:

Design Requirements:

1. Must reference:
 - Disparate metrics (e.g., departmental budgets vs regional sales)
 - Chronological differences (quarterly/annual comparisons)
 - Contrasting categories (actual vs projected figures)
2. Dependency Rules:
 - Each segment provides unique essential information
 - No overlapping data between required sections
3. Answerability Constraints:
 - Unanswerable without the provided context
 - Answer must demonstrate relational understanding
 - Require explicit mention of both referenced sections
4. Response Specifications:
 - Maximum answer length: avoid too long
 - Comparisons must use contextually appropriate units

Example QAs:

{Q: “How did the average revenue per Full Course Equivalent (FCE) enrollment change in the Degree Program Segment compared to the Alternative Credential Segment from 2022 to 2023?”, A: “The average revenue per FCE enrollment increased by 17.8% in the Degree Program Segment from \$2,447 in 2022 to \$2,883 in 2023, while it decreased by 6% in the Alternative Credential Segment from \$3,897 in 2022 to \$3,662 in 2023.”}

...

Financial Statement: {Context}

Figure 9. The prompt for generating QA pairs.

Example prompt used for generating *Hallucination Detection* QAs in financial statements

Task: You are a financial analyst. Design a pseudo-relevant question that appears answerable but actually lacks sufficient basis in the financial statement, following these principles:

Design Requirements:

1. Surface-level Relevance:
 - Use document-specific terminology
 - Reference actual sections/metrics as distractors
2. Unanswerability Guarantees:
 - Absolutely not mentioned in the context
 - Missing critical data points required for resolution
 - No inferential path from provided information
3. Confirm absence of:
 - Direct mentions
 - Implied values
 - Comparable proxies

Example QAs:

{Q: “What measures are being taken to mitigate foreign currency risk?”, A: “The document does not specify any measures being taken to mitigate foreign currency risk.”}

{Q: “What is the company’s market share in the non-combustible nicotine-related products sector?”, A: “The document does not provide a specific percentage for the company’s market share in the non-combustible nicotine-related products sector.”}

...

Financial Statement: {Context}

Figure 10. The prompt for generating QA pairs.

Examples of *location* task

Novel

[Query] What does Mrs. Fitzgerald think of Violet and her family?

[Ground-truth Answer] She believes they are very common.

=====

Academic Paper

[Query] According to paper 2, what impact does self-critiquing have on the plan generation performance of LLMs in comparison to using an external verifier?

[Ground-truth Answer] Self-critiquing degrades the plan generation performance compared to using an external, sound verifier.

=====

Financial Statements

[Query] How much cash was provided by financing activities during the three months ended March 31, 2024?

[Ground-truth Answer] \$6.3 million.

Figure 11. The examples of location task.

Examples of *reasoning* task

Novel

[Query] Why did Richard know by the second meeting that Karen didn't like sweets?

[Ground-truth Answer] Because during their first meeting over a meal, Richard noticed that Karen quietly disposed of the dessert on her plate.

=====

Academic Paper

[Query] Based on paper 2, why does the LLM+LLM backprompting system underperform compared to the LLM+VAL system in plan generation?

[Ground-truth Answer] The LLM+LLM backprompting system underperforms because the verifier LLM produces a significant number of false positives, declaring incorrect plans as valid, which undermines the reliability of the system. In contrast, the LLM+VAL system uses an external sound verifier, VAL, which provides accurate validation, leading to better overall performance due to fewer verification errors.

=====

Financial Statements

[Query] What percentage of the total fair value of the liabilities assumed during the merger with DHC were warrant liabilities?

[Ground-truth Answer] To calculate the percentage of warrant liabilities in the merger with DHC, sum the warrant liability's value, \$1,913,737, and divide it by the total net liabilities assumed, \$9,863,196. The percentage is found by $(1,913,737 / 9,863,196) * 100$, which equals approximately 19.4%.

Figure 12. The examples of reasoning task.

Examples of *comparison* task

Novel

[Query] How does the narrator’s attitude towards Turkey and Nippers compare to his attitude towards Bartleby?

[Ground-truth Answer] The narrator is willing to overlook Turkey’s and Nippers’ flaws due to their usefulness, just as he tolerates Bartleby’s peculiar behavior because of his steady work and presence.

=====

Academic Paper

[Query] How do the approaches for incorporating context into ranking functions differ between the method proposed in paper 0 and in paper 1?

[Ground-truth Answer] Paper 0 incorporates context using delta features comparing neighboring items, while paper 1 uses a self-attention mechanism to account for interactions between items during both training and inference.

=====

Financial Statements

[Query] How did the total current assets and total current liabilities change for CISO Global, Inc. from December 31, 2023, to March 31, 2024?

[Ground-truth Answer] CISO Global’s total current assets decreased from \$10,957,814 on December 31, 2023, to \$9,276,063 on March 31, 2024, while total current liabilities increased from \$26,071,102 to \$32,604,126 in the same period.

Figure 13. The examples of comparison task.

Examples of *hallucination detection* task

Novel

[Query] What type of flower did Alexander place on Violet’s grave in the cemetery?

[Ground-truth Answer] TThe text does not mention Alexander placing any type of flower on Violet’s grave in the cemetery.

=====

Academic Paper

[Query] In paper 2, what are the implications of AI-enhanced NMR processing on the prediction of chemical reaction pathways?

[Ground-truth Answer] Paper 2 does not discuss the implications of AI-enhanced NMR processing on the prediction of chemical reaction pathways.

=====

Financial Statements

[Query] What were the results of the company’s environmental sustainability initiatives?

[Ground-truth Answer] The financial statement does not mention or provide any details about the results of the company’s environmental sustainability initiatives.

Figure 14. The examples of hallucination detection task.