

Where am I? Large Language Models Wandering between Semantics and Structures in Long Contexts

Seonmin Koo^{1*}, Jinsung Kim^{1*}, YoungJoon Jang¹, Chanjun Park^{2†}, Heuseok Lim^{1†}

¹Department of Computer science and Engineering, Korea University

²Upstage AI

{fhdahd, jin62304, limhseok}@korea.ac.kr

yjoonjang34@gmail.com

chanjun.park@upstage.ai

Abstract

As the utilization of Large Language Models (LLMs) becomes more widespread, there is a growing demand for their ability to handle more complex and longer external knowledge across various use cases. Most existing evaluations of the open-ended question answering (ODQA) task, which necessitates the use of external knowledge, focus solely on whether the model provides the correct answer. However, even when LLMs answer correctly, they often fail to provide an obvious source for their responses. Therefore, it is necessary to jointly evaluate and verify the correctness of the answers and the appropriateness of grounded evidence in complex external contexts. To address this issue, we examine the phenomenon of discrepancies in abilities across two distinct tasks—QA and evidence selection—when performed simultaneously, from the perspective of task alignment. To verify LLMs’ task alignment, we introduce a verification framework and resources considering both semantic relevancy and structural diversity of the given long context knowledge. Through extensive experiments and detailed analysis, we provide insights into the task misalignment between QA and evidence selection. Our code and resources can be found at <https://github.com/seonminkoo/WAI>.

1 Introduction

Open-domain question answering (ODQA) task aims to generate accurate answers for a wide range of contexts and queries (Chen et al., 2017; Yang et al., 2019). In the ODQA task, large language models (LLMs) should be capable of leveraging fluent generative abilities and retrieval-augmented generation (RAG)-based capabilities to effectively utilize more complicated and longer external knowledge from the real world.

* Equally contributed.

† Corresponding author.

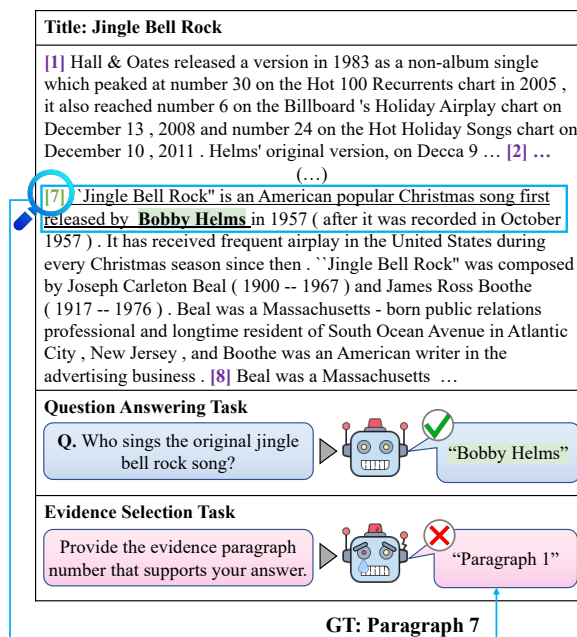


Figure 1: Example of task alignment verification. LLM correctly answers the given question but selects an incorrect evidence paragraph that does not support the answer. We find that LLMs often exhibit a task misalignment between QA and evidence selection.

One of the challenges in leveraging diverse real-world knowledge is that knowledge contexts may not contain only helpful information for generation (Baek et al., 2023). In other words, unrelated or fascinating but partially relevant information that may confuse the model can be spread across various locations within the knowledge context (Ji et al., 2023; Bang et al., 2023; Yang et al., 2023). This can lead LLMs to generate hallucinated content that contradicts reality or fail to clearly present the evidence source for their responses, as illustrated in Figure 1. In particular, this phenomenon becomes more severe as the knowledge context becomes longer (Huang et al., 2023).

Despite these challenges, conventional QA research primarily evaluates whether model outputs

are included in the gold answer using metrics such as Exact Match (EM) score, thus often not focusing on the correctness of the sources underlying the generated responses (Chen et al., 2017; Lewis et al., 2020; Izacard and Grave, 2020). Recent studies on LLMs’ evidence selection ability have explored various approaches, including modeling and prompting (Jo et al., 2021; Wei et al., 2022; Chen et al., 2022; Wang et al., 2023; Kong et al., 2023). However, these studies have typically treated QA and evidence source selection as separate tasks, relatively overlooking the discrepancy in the LLMs’ task alignment capabilities.

Bridging this discrepancy is crucial for ensuring the explainability of LLMs’ outputs and enhancing end-user reliability (Nakano et al., 2021; Menick et al., 2022; Ras et al., 2022). Therefore, when LLMs handle extensive external knowledge, it is essential to sufficiently verify not only the accuracy of responses but also the appropriateness of the corresponding evidence selected by models.

To address these issues, we conduct an integrated and systematic study on the alignment between answer correctness and grounding appropriateness, which have traditionally been studied separately, in challenging long context settings. We explore task alignment by simultaneously performing two distinct tasks—QA and evidence selection, utilizing the same features.

For systematic validation, we conduct an empirical study on LLMs’ alignment capabilities, considering both the semantic relevancy and structural diversity of the long context knowledge they utilize. Specifically, we construct semantically diverse surrounding input contexts by retrieving relevant and irrelevant information from multiple documents using the RAG approach. Additionally, we investigate the LLMs’ behavior in detail by structurally diversifying the position of ground-truth evidence within complex contexts.

From a task alignment perspective, we provide insights into the discrepancy between LLMs’ capabilities in QA and evidence selection through extensive experiments and detailed analysis. Furthermore, we conduct additional experiments and analyses based on context lengths and prompting strategies.

Our contributions are threefold: **i)** We introduce a framework and resource construction method for validating task misalignment occurring in real-world long context-based generation scenarios. **ii)** To achieve systematic validation, we devise the con-

cepts of semantic relevancy and structural diversity, effectively investigating LLMs’ task alignment capabilities and achieving enhanced explainability. **iii)** Through extensive experiments and analysis, we provide various insights on the LLMs’ behavior in QA and evidence selection tasks when generating based on long contexts.

2 Related Works

2.1 Long context LLMs

To understand how LLMs handle long context inputs, (Li et al., 2024a; Levy et al., 2024) proposed new datasets and discovered that LLMs exhibit notable degradation in reasoning performance at much shorter input lengths than their technical maximum. Similarly, Liu et al. (2023) demonstrated the “lost-in-the-middle” phenomenon, where LLMs show lower accuracy when evidence for a query is located in the middle of a long context input. (Junqing et al., 2023; Zhang et al., 2024) proposed methodologies to mitigate these issues.

To evaluate the attributions of LLM-generated text, ZeroScrolls (Shaham et al., 2023), LongBench (Bai et al., 2023), Cab (Zhang et al., 2023), LooGLE (Li et al., 2023), and ELITR-Bench (Thonet et al., 2024) proposed to evaluation method of LLMs on long context inputs. Additionally, L-Eval (An et al., 2023) constructed datasets and investigated the effectiveness of evaluation metrics for standardized assessment.

2.2 LLMs with Attribution Retrieval/Generation

To mitigate hallucination in LLMs, several methodologies have been developed to generate or retrieve attributions along with sentence generation (Menick et al., 2022; Gao et al., 2023; Jain et al., 2023; Huo et al., 2023). These approaches help ensure that the responses generated by LLMs are based on reliable sources. In the retrieve-and-read paradigm, a retrieval for relevant documents within a corpus, and the language model reorders them to extract an answer (Chen et al., 2017; Zhu et al., 2021). Generative retrieval has emerged as an alternative to the traditional retrieve-and-read pipeline, integrating search and generation into a single process. This approach controls the generation process by ensuring that the LLM’s responses include keywords or content from the retrieved information (De Cao et al., 2020; Metzler et al., 2021; Li et al., 2024b; Sun et al., 2024).

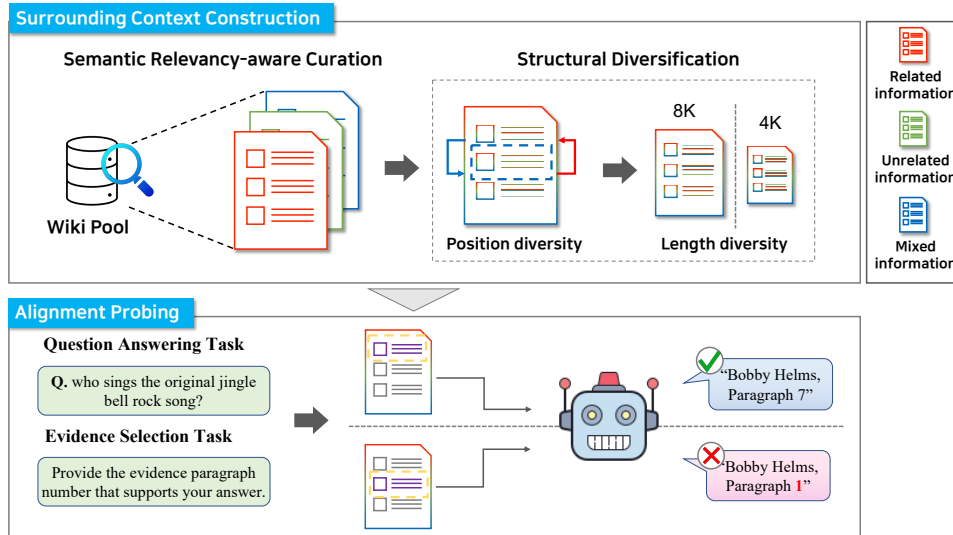


Figure 2: Overview of our task alignment verification. The upper box illustrates the construction of the surrounding context in real-world environments (related, unrelated, and mixed information) through semantic relevancy and structural diversification. The lower box demonstrates alignment probing in QA and evidence selection tasks. LLMs may exhibit task alignment discrepancies for the same query.

3 Experimental Setup

In this section, we describe the process of constructing resources to verify the alignment between answer correctness and appropriateness of evidence selection in LLMs’ responses within long-context scenarios. Additionally, we explain the models and evaluation metrics adopted for the tasks.

3.1 Datasets

Inspired by previous studies that validate the capabilities of existing LLMs (Chen et al., 2023; Ma et al., 2023), we experiment with the task alignment of LLMs in long contexts through QA and evidence selection tasks. For our experiment, we utilize the NQ dataset (Kwiatkowski et al., 2019), which is constructed based on Wikipedia that serves as a realistic world knowledge pool similar to a commercial search engine. This dataset includes questions formulated by users of search engines and answers manually extracted from Wikipedia. NQ dataset contains questions of formats such as ‘who’, ‘when’, and ‘how many’. Detailed statistics of the dataset can be found in Appendix A.1.

To set up the challenging long-context-based tasks, we extract contexts exceeding 16k from various Wikipedia sources. For the diverse types of questions and context included, the model is expected to provide both the answer and the evidence paragraph.

3.2 Surrounding Context Construction

LLMs should be able to effectively utilize various external knowledge sources provided through methods such as RAG. However, in the real world, knowledge context may not contain information only beneficial for task performance, and attractive distractors could be spread across various sources (Ji et al., 2023; Bang et al., 2023; Yang et al., 2023). Therefore, considering these real-world challenges, we construct the surrounding contexts as illustrated in the upper part of Figure 2, considering both semantic relevancy (§ 3.2.1) and structural diversity (§ 3.2.2).

3.2.1 Semantic Relevancy-aware Curation

Related Information Related information refers to contents with highly relevant to the query. To prepare for the 16k setting, we first select 100 documents over 16k in length. Each document is embedded using OpenAI’s text-embedding-3-large model (OpenAI-Blog, 2022) and stored in Pinecone¹ (please refer to Appendix A.2). From the stored documents, we extract paragraphs with high cosine similarity to the query, reconstructing a context set based on the related information.

Unlike previous studies that simply concatenate multiple short documents to create contexts, we build a more semantics-aware context by extracting and ranking relevant paragraphs from a single

¹<https://www.pinecone.io>

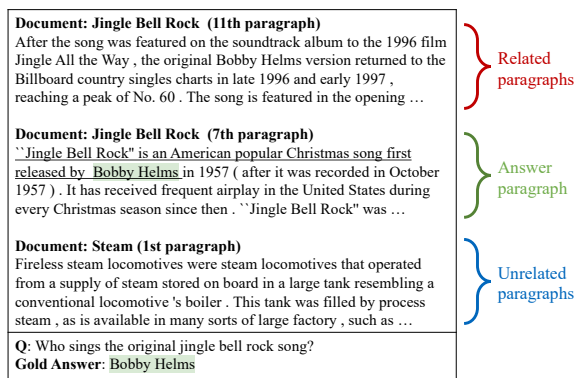


Figure 3: An example of our mixed context construction result. Context contains related or unrelated paragraphs along with the answer paragraph.

document. We utilize the contexts reconstructed with related information to verify whether they positively contribute to task performance or act as misleading evidence that may adversely affect the model.

Unrelated Information Unrelated information refers to contents with low relevance to the query. Most existing research extracts unrelated information by randomly selecting documents from the search pool (Yoran et al., 2023; Cuconasu et al., 2024). However, the random sampling method may include paragraphs related to the query. To prevent this, we apply low-rank retrieval. Specifically, we calculate the similarity between the query and the documents stored in Pinecone, then select the document with the lowest relevance.

Subsequently, we build an unrelated context by sequentially adding paragraphs that have low cosine similarity with the query from the selected unrelated documents. This approach allows us to construct a semantic relevancy-aware setting that is strictly differentiated from the related information.

Mixed information Mixed information refers to contexts containing both related and unrelated information. In real-world scenarios, considering that queries rarely contain fully related or unrelated information alone, we reconstruct a more challenging setting of mixed context for the model.

We build the context to include an equal proportion of the previously extracted related and unrelated information based on their similarity to the query. From the related information, we sequentially select the top paragraphs with the highest cosine similarity to the query. Conversely, from the unrelated information, we sequentially select the

bottom paragraphs with the lowest similarity to the query. This approach enables us to verify whether the model can provide accurate responses by effectively filtering the task-relevant information within semantically mixed contexts.

Figure 3 illustrates an example of the mixed context from our proposed data construction method. The original example includes a document titled "Jingle Bell Rock," a question-answer pair, and the ground truth (GT) paragraph supporting the answer. The related information comprises paragraphs highly relevant to the query. Consequently, within the same document titled "Jingle Bell Rock," highly relevant paragraphs are selected. In contrast, the unrelated information consists of the least relevant paragraphs from the least relevant document. Therefore, text from an unrelated document titled "Steam" is selected. This mixed composition approach is useful for evaluating whether the model can clearly distinguish between relevant and irrelevant information.

3.2.2 Structural Diversification

Position Diversity We consider 'position diversity' when performing QA and evidence selection tasks, as irrelevant answers or incorrect but attractive answers that do not directly substantiate the task can be spread across various positions within the given context. The 'Structural diversification' step in the upper box of Figure 2 demonstrates the process wherein the position between the ground-truth paragraph and surrounding paragraphs variably shifts within a context constructed upon diverse semantic relevancy.

Specifically, we diversify the position of the GT evidence paragraph within the context into three cases: at the top, middle, and bottom. For instance, if the sequence of paragraphs in the initial context is 'related information-answer paragraph-unrelated information,' we vary this into forms such as 'answer-related-unrelated,' 'unrelated-answer-related,' and 'unrelated-related-answer.' Through this approach, we can validate task alignment in a more challenging and diversified setting.

Length Diversity In addition to position diversity, we build contexts of 8k and 4k lengths alongside 16k contexts to compare the model's behavior with different context lengths. The 8k and 4k contexts are also constructed based on semantic relevancy, following the same process described

Models	F1	Rouge-L	EM	F1	Rouge-L	EM	F1	Rouge-L	EM
	<i>Related Information</i>								
	RAR			ARR			RRA		
GPT-3.5	0.16	0.14	0.46	0.19	0.17	0.42	0.16	0.14	0.43
Mixtral	0.13	0.11	0.53	0.12	0.11	0.53	0.14	0.13	0.57
Claude3	0.14	0.13	0.56	0.16	0.14	0.55	0.14	0.13	0.54
	<i>Unrelated Information</i>								
	UAU			AUU			UUA		
GPT-3.5	0.06	0.06	0.13	0.16	0.13	0.34	0.09	0.08	0.13
Mixtral	0.18	0.16	0.59	0.18	0.17	0.64	0.14	0.13	0.61
Claude3	0.10	0.09	0.31	0.12	0.11	0.45	0.08	0.07	0.19
	<i>Mixed information</i>								
	MAM			AMM			MMA		
GPT-3.5	0.13	0.12	0.34	0.16	0.14	0.42	0.15	0.13	0.31
Mixtral	0.13	0.12	0.54	0.15	0.13	0.57	0.14	0.12	0.58
Claude3	0.14	0.12	0.55	0.15	0.14	0.60	0.16	0.13	0.57

Table 1: Evaluation results of QA task. **R**, **A**, and **U** denote paragraphs containing related information, the answer, and unrelated information, respectively. **M** indicates a mixture of related and unrelated information. The arrangement illustrates the semantic and structural composition of the context. For example, **RAR** indicates the context sequence ‘related-answer-related’.

in Section 3.2.1, and include paragraphs with related and unrelated information. Furthermore, position diversification is applied to the constructed 8k and 4k contexts. This involves shifting the position of the GT paragraph within the context to various positions, such as top, middle, and bottom, to evaluate the model’s performance under different conditions.

3.3 Models

For all experiments, we adopted the GPT-3.5-turbo-16k model using OpenAI’s ChatGPT API (OpenAI-Blog, 2022). The hyperparameters were set to their default values as recommended by OpenAI. Additionally, to handle various types of LLMs, we also employed Claude 3 (Anthropic, 2024) and the Mixtral-8x7B model. Specifically, we used the claude-3-haiku-20240307 version for Claude 3.

Notably, it is important to highlight that ChatGPT or other models might occasionally generate empty responses due to network transmission timeouts or API overload. In such cases, we followed the standard practice of resubmitting the request until obtaining non-empty responses. To prevent any potential influence from prior responses, we cleared the conversation history each time and submitted a new query to the model.

3.4 Evaluation Metrics

To verify the task alignment of LLMs, we jointly perform QA and evidence selection tasks. An exam-

ple of the prompt template for the tasks is provided in Appendix A.3.

3.4.1 Question Answering Task

To evaluate the quality of the generated answers, we use the F1-score (Sokolova et al., 2006) as the primary metric, which is the harmonic mean of precision and recall. Additionally, we employ ROUGE-L (Lin, 2004), based on n-gram matching with the reference, to measure the similarity between the generated answers and the reference answers. We also utilize the Exact Match (EM) score as supplementary indicator to assess the accuracy of generated answers.

3.4.2 Evidence Selection Task

To evaluate the preciseness of selecting evidence paragraphs corresponding to answers to queries, we use Accuracy (%). We extract evidence numbers from model-generated responses and compare them with GT evidence paragraph numbers.

In preliminary experiments, we observed instances where LLMs did not respond to the evidence selection task. Therefore, we also measure the rate of when the model does not provide an answer to the instruction (no answer). If the generated response does not include an evidence number, it is considered a ‘no answer’ case.

4 Experiments and Analysis

In this section, we analyze the results of experimenting with LLMs by feeding challenging contexts that are semantically and structurally diversified in Section 3.2. Additionally, we provide various qualitative analyses (refer to Appendix B).

4.1 How Do Context Changes in Semantic and Structure Affect LLMs’ QA Abilities?

Following previous research on LLMs responses to context (Xu et al., 2023; Jiang et al., 2023; Liu et al., 2024), we verify the answer generation abilities of LLMs through the QA task. Table 1 shows the evaluation results of the QA task. We observe the capabilities of LLMs in terms of semantic and structural changes across various combinations of diverse information types: related (R), unrelated (U), and mixed information (M). Abbreviations such as ‘RAR’ represent the order of context composition. For example, ‘RAR’ indicates the sequence ‘related-answer-related’.

First, the results concerning *position diversity* are as follows: When the surrounding contexts are composed only of related information, i.e., RAR, ARR, and RRA, the F1 performance is lower when the answer paragraph is positioned in the middle compared to the other two cases. This result aligns with previous findings that LLMs tend to lose information positioned in the middle of the context (Liu et al., 2023). Similarly, in contexts composed of unrelated and mixed information, the F1 performance is generally lower when the answer paragraph positioned in middle—UAU and MAM—compared to other positions.

Moreover, the results concerning changes in *semantic relevancy* during context composition are as follows: When the surrounding contexts are composed solely of related information, the performance is highest regardless of the answer paragraph’s position compared to contexts with unrelated or mixed information. Conversely, the lowest performance is observed when only unrelated information is provided.

Therefore, semantically, surrounding contexts composed solely of related information positively impact LLMs’ QA abilities. Structurally, the answer paragraph’s position at the top of the context has the most positive effect. We provide insights into the QA capabilities of LLMs across various contexts by applying semantic and structural concepts, thereby enhancing their interpretability.

Models	Acc.		no ans.		Acc.		no ans.	
	Related Information							
	RAR		ARR		RRA			
GPT-3.5	0.62	0.10	0.62	0.09	0.50	0.12		
Mixtral	0.62	0.04	0.48	0.05	0.60	0.06		
Claude3	0.59	0.01	0.62	0.01	0.45	0.01		
Avg.	0.61	0.05	0.57	0.05	0.52	0.06		
Unrelated Information								
	UAU		AUU		UUA			
GPT-3.5	0.56	0.44	0.72	0.24	0.52	0.46		
Mixtral	0.95	0.04	0.95	0.05	0.90	0.05		
Claude3	0.53	0.47	0.67	0.33	0.38	0.60		
Avg.	0.68	0.32	0.78	0.21	0.60	0.37		
Mixed Information								
	MAM		AMM		MMA			
GPT-3.5	0.45	0.22	0.52	0.20	0.51	0.21		
Mixtral	0.48	0.01	0.55	0.02	0.63	0.04		
Claude3	0.57	0.03	0.79	0.01	0.60	0.01		
Avg.	0.50	0.09	0.62	0.08	0.58	0.09		

Table 2: Evaluation results of evidence selection task. Avg., Acc. and no ans. indicate the average of the scores, accuracy and, no answer, respectively.

4.2 How Do Context Changes in Semantic and Structure Affect LLMs’ Evidence Selection Abilities?

In real-world scenarios, LLMs’ ability to clearly present the evidence source amidst confusing information for their generated responses is crucial. Table 2 shows the evaluation results of the evidence selection task.

Firstly, from a semantic perspective, LLMs achieve the highest accuracy when the surrounding contexts are composed solely of unrelated information. Conversely, the evidence selection ability of LLMs is most diminished when the contexts are composed of related information. Moreover, from a structural perspective, the evidence selection capability of LLMs is observed to be independent of the position of answer paragraphs.

The proportion of ‘no answer’ responses based on the location of the answer paragraph is in the following order: in the bottom, middle, and top, respectively. In other words, as the answer paragraph is placed later in the context, LLMs’ instruction-following ability for the evidence selection task decreases significantly.

Therefore, we find that LLMs’ evidence selection ability improves when the context is composed solely of unrelated information and that the position of the answer paragraph does not significantly affect LLMs’ performance.

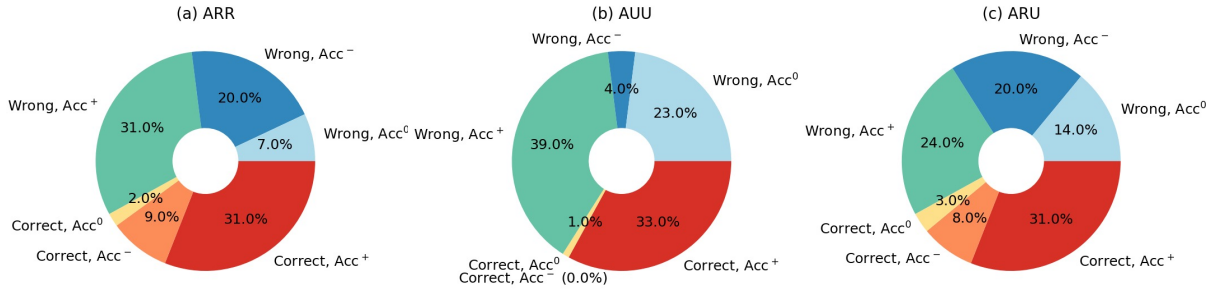


Figure 4: Alignment of QA and evidence selection tasks. Each of the six sectors of the chart corresponds to a QA task’s performance (correct, incorrect) and evidence selection ability (Acc^+ , Acc^- , Acc^0).

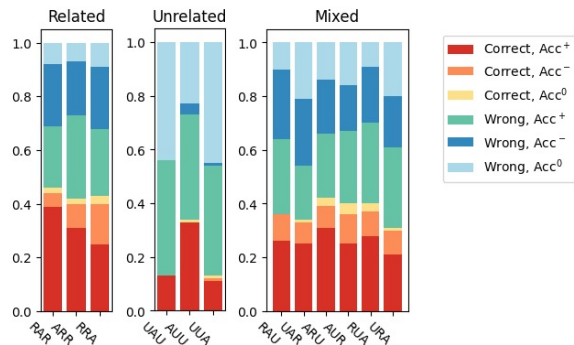


Figure 5: Alignment of QA and evidence selection tasks for all variations. The height of each bar represents the accuracy of evidence extraction for QA. We provide detailed mixed settings, such as ‘MAM \rightarrow UAR and RAU’.

4.3 Alignment Probing: Do LLMs’ QA Capabilities Align with Evidence Selection Abilities?

As seen in the lower box of Figure 2, LLMs may fail to correctly identify the evidence source from the context even if they provide the correct answer. Bridging the misalignment between LLM’s abilities across tasks is crucial for preventing hallucinations and ensuring the explainability of LLM’s generative behavior. To this end, we explore the task alignment between LLMs’ answer generation and evidence selection abilities.

Figure 4 and Figure 5 illustrate the alignment of the QA task and the evidence selection task. The terminology used in the figures is categorized as follows:

- The correctness of answers of LLMs in QA tasks is indicated as either ‘Correct’ or ‘Wrong.’
- The accuracy of evidence selection of LLMs is indicated as correct (Acc^+), incorrect (Acc^-),

or no answer (Acc^0).

We find discrepancies in the alignment between QA and evidence selection tasks. In Figure 4, the proportion of cases where both QA and evidence selection tasks are performed correctly (Correct, Acc^+ in red) is comparable to the proportion where QA is incorrect but evidence selection is correct (Wrong, Acc^+ in green). Additionally, the proportion of cases where both QA and evidence selection tasks are performed incorrectly (Wrong, Acc^- in blue) is low.

This indicates that the LLMs’ abilities in the QA task do not necessarily lead to their evidence selection abilities, resulting in the task misalignment phenomenon. Particularly, in Figure 5, the proportion of cases where both QA and evidence selection tasks are incorrect (Wrong, Acc^- in blue) is lowest in the unrelated information context compared to the related or mixed information contexts. This suggests that task misalignment is most pronounced in contexts involving unrelated contexts.

Therefore, we provide insight that even when LLMs are given contexts with the same semantic and structural context, performance mismatches between the tasks can occur. For this reason, unrelated information can act as noise and interfere with generation in the QA tasks where the LLMs’ ability to synthesize information and evidence is crucial. Conversely, in the evidence selection task, related information can confuse the model when extracting the evidence source.

4.4 How Does Context Length Affect LLMs’ Task Alignment Abilities?

Figure 6 and Figure 7 demonstrate the performance on QA tasks and evidence selection tasks across various context lengths, respectively. In QA tasks,

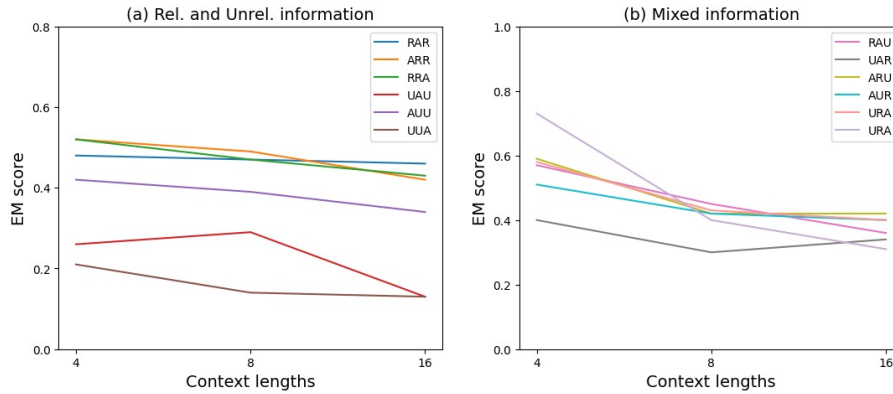


Figure 6: Evaluation results of QA task by context length (4k, 8k, 16k). ‘Rel.’ and ‘Unrel.’ denote related and unrelated contexts, respectively.

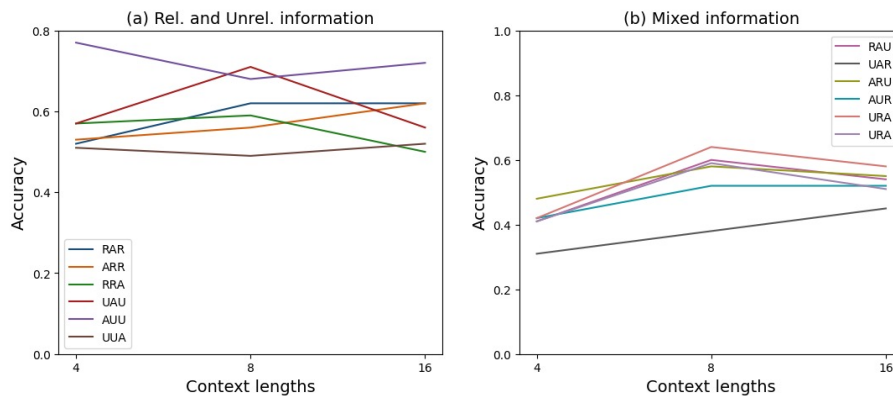


Figure 7: Evaluation results of evidence selection task by context length (4k, 8k, 16k). ‘Rel.’ and ‘Unrel.’ denote related and unrelated contexts, respectively.

performance tends to decrease as the length of related or unrelated information provided increases.

In contrast, in the evidence selection task, performance can improve, such as in the RAR case, or show different score change tendencies compared to the QA task. Additionally, in mixed information contexts, the performance changes with respect to length are opposite for the QA and evidence selection tasks. Most cases show lower QA abilities at 8k or 16k lengths, while the evidence source identification ability is lowest at 4k.

This suggests that the task misalignment is most pronounced at the 8k length. Therefore, we provide insights into a discrepancy in LLMs’ performance tendency regarding the length of contexts between the QA and evidence selection tasks.

4.5 Do Prompting Strategies Mitigate LLM’s Task Misalignment?

Our efforts to quantify and verify task misalignment in LLMs, considering semantic relevancy and

structural diversity, can be the significant foundation for mitigating the discrepancy.

As part of the exploration of mitigation feasibility, we conduct several prompting experiments that may offer extended insights and understanding. According to Figure 5, task misalignment was most pronounced in the RRA (Correct, ACC^-) and UAU (Wrong, ACC^+) settings. Therefore, we investigate whether applying several prompting strategies to these two settings can mitigate LLM’s task misalignment. In particular, Chain-of-Thought (CoT) (Wei et al., 2022) and In-Context Learning (ICL) (Brown, 2020) methods are explored, which are the most widely used to induce in-depth reasoning. For the ICL method, a 1-shot exemplar in the form of (context, query, answer) is provided.

Table 4 and Table 5 show the experimental results in the individual tasks, QA and evidence selection, respectively. First, the performance of QA demonstrates a general upward tendency when ad-

	<i>Misalignment Case</i>		<i>Others</i>			
	Correct, Acc ⁻	Wrong, Acc ⁺	Correct, Acc ⁺	Correct, Acc ⁰	Wrong, Acc ⁻	Wrong, Acc ⁰
RRA						
Baseline	0.15	0.25	0.25	0.03	0.23	0.09
CoT	0.21 (+0.06)	0.23 (-0.02)	0.23	0.02	0.27	0.04
ICL	0.16 (+0.01)	0.15 (-0.10)	0.16	0.09	0.17	0.27
UAU						
Baseline	0.00	0.43	0.13	0.00	0.00	0.44
CoT	0.00 (-)	0.33 (-0.10)	0.39	0.03	0.04	0.21
ICL	0.00 (-)	0.22 (-0.21)	0.14	0.07	0.00	0.57

Table 3: Experimental results on prompting strategies for mitigating task misalignment. “Misalignment Case” (left) indicates where the model responds differently to each task, QA and evidence selection. “Others” (right) represents aligned or non-response cases.

	F1	Rouge-L	EM	F1	Rouge-L	EM
	RRA			UAU		
Baseline	0.16	0.14	0.43	0.06	0.06	0.13
CoT	0.17	0.15	0.46	0.16	0.13	0.42
ICL	0.24	0.24	0.41	0.17	0.17	0.21

Table 4: Experimental results on prompting strategies for QA task.

vanced prompting strategies are adopted, as shown in Table 4. On the other hand, only the CoT method improves accuracy, and no answer rate in the evidence selection task, and the ICL method generally deteriorates both performances, as shown in Table 5.

Notably, the task alignment performance in Table 3 shows inconsistent distribution according to the composition setting of context, such as RRA and UAU. For instance, when CoT prompting is applied in the RRA setting, the misalignment rate of *Correct, Acc⁻* case increases from 0.15 to 0.21, indicating a decline in task alignment capability. On the other hand, in the UAU setting, the task alignment capability improves from 0.43 to 0.33 (CoT) and 0.22 (ICL), respectively, in the case of *Wrong, Acc⁺*.

These results suggest that while applying prompting strategies sometimes results in improvements in individual tasks, it may not necessarily mitigate LLMs’ task misalignment. Therefore, it is crucial to specifically design an optimized approach to bridge this discrepancy in LLMs’ task alignment capabilities, ensuring the explainability of their outputs and enhancing end-user reliability. In future work, researchers could extend this study by exploring methods to enhance the model’s task alignment ability in long-context situations.

	Acc.	no ans. (↓)	Acc.	no ans. (↓)
	RRA		UAU	
Baseline	0.50	0.12	0.56	0.44
CoT	0.46	0.06	0.72	0.24
ICL	0.31	0.36	0.36	0.64

Table 5: Experimental results on prompting strategies for evidence selection task.

5 Conclusion

One of the challenges of utilizing long-context knowledge in real-world scenarios is that the knowledge context does not contain information only helpful for generation. Consequently, LLMs often generate hallucinated contents or fail to adequately present the evidence source for their responses. Therefore, we examine and interpret the phenomenon of discrepancies in LLMs’ abilities when performing two different tasks—QA and evidence selection—simultaneously using the same input features, from the perspective of *task alignment*. To address this, we propose a task alignment verification framework and resource construction methodology based on the concepts of semantic relevance and structural diversity. Through extensive experiments and analysis in settings that closely resemble real-world scenarios, we provide various insights to bridge the task alignment gap observed in LLMs. We hope our work serves as useful reference material for preventing hallucinations and enhancing the explainability of LLMs’ generative behavior when given long contexts.

Limitations

Despite the significant contributions of this study, there are several limitations that need to be ac-

knowledge.

First, due to resource constraints, our experiments were limited to only three different LLMs. While these models were selected to provide a diverse representation, a broader range of models could potentially offer more comprehensive insights into task alignment issues. Future work should consider including a wider variety of models to validate and extend the findings of this study.

Second, the constructed contexts in our experiments were designed to reflect realistic scenarios as much as possible, but they still may not capture the full complexity of real-world applications. The dynamics of user interactions and the variety of external knowledge sources in practical settings might present additional challenges not fully addressed by our study. Therefore, real-world testing and validation in diverse application domains would be beneficial to further substantiate our findings.

Finally, while our framework and methodology provide a robust approach to evaluating task alignment, there is room for refinement in terms of measurement granularity and the evaluation metrics used. More nuanced metrics could potentially offer deeper insights into the alignment discrepancies observed between QA and evidence selection tasks.

Despite these limitations, this study offers valuable contributions to understanding the alignment between QA and evidence selection tasks in LLMs, particularly in the context of long and complex external knowledge. Our findings highlight important areas for improvement and provide a foundation for future research aimed at enhancing the reliability and explanatory power of LLM-generated responses in real-world applications.

Ethics Statement

We discuss the main ethical considerations of the model we proposed: (1) Privacy. the datasets adopted to construct our model provide factual knowledge and fictional person’s preferences, and our model does not contain privacy issues. (2) Potential problems. Although we take conscientious steps to ensure the quality of our models, there can still be potential problems with the generated responses’ quality, which can lead to incorrect predictions in applications that leverage factual information and human preferences. (3) Model deployment. Our approach employs the Large Language Models (LLMs) for the downstream tasks, which have the risk of reflecting the bias of the training data. It is

a well-known threat in tasks using LLMs, and we should be careful about social impact when using this method since our model aims to handle factual knowledge.

Acknowledgement

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI). This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425). This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. RS-2022-II220369, 2022-0-00369 (Part 4) Development of AI Technology to support Expert Decision-making that can Explain the Reasons/Grounds for Judgment Results based on Expert Knowledge)

References

- Chenxin An, Shansan Gong, Ming Zhong, 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*.
- AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.
- Jinheon Baek, Alham Fikri Aji, and Amir Saffari. 2023. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. *arXiv preprint arXiv:2306.04136*.
- Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.
- Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multi-task, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.
- Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

- Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.
- Jiangjie Chen, Wei Shi, Ziquan Fu, Sijie Cheng, Lei Li, and Yanghua Xiao. 2023. Say what you mean! large language models speak too positively about negative commonsense knowledge. *arXiv preprint arXiv:2305.05976*.
- Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.
- Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. *arXiv preprint arXiv:2401.14887*.
- Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive entity retrieval. *arXiv preprint arXiv:2010.00904*.
- Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.
- Siqing Huo, Negar Arabzadeh, and Charles Clarke. 2023. Retrieving supporting evidence for generative question answering. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 11–20.
- Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.
- Palak Jain, Livio Soares, and Tom Kwiatkowski. 2023. 1-pager: One pass answer generation and evidence retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14529–14543.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.
- Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. *arXiv preprint arXiv:2310.06839*.
- Yohan Jo, Haneul Yoo, JinYeong Bak, Alice Oh, Chris Reed, and Eduard Hovy. 2021. Knowledge-enhanced evidence retrieval for counterargument generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3074–3094, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- He Junqing, Pan Kunhao, Dong Xiaoqun, Song Zhuoyang, Liu Yibo, Liang Yuxin, Wang Hao, Sun Qianguo, Zhang Songxin, Xie Zejian, et al. 2023. Never lost in the middle: Improving large language models via attention strengthening question answering. *arXiv preprint arXiv:2311.09198*.
- Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xin Zhou. 2023. Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702*.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. *Natural questions: A benchmark for question answering research*. *Transactions of the Association for Computational Linguistics*, 7:452–466.
- 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*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2023. Loogle: Can long-context language models understand long contexts? *arXiv preprint arXiv:2311.04939*.
- Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024a. Long-context llms struggle with long in-context learning. *arXiv preprint arXiv:2404.02060*.
- Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. 2024b. Learning to rank in generative retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8716–8723.
- Chin-Yew Lin. 2004. *ROUGE: A package for automatic evaluation of summaries*. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

- Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *corr abs/2307.03172* (2023).
- 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.
- Xilai Ma, Jing Li, and Min Zhang. 2023. [Chain of thought with explicit evidence reasoning for few-shot relation extraction](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2334–2352, Singapore. Association for Computational Linguistics.
- Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.
- Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. 2021. Rethinking search: making domain experts out of dilettantes. In *Acm sigir forum*, volume 55, pages 1–27. ACM New York, NY, USA.
- Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.
- OpenAI-Blog. 2022. [Chatgpt: Optimizing language models for dialogue](#).
- Gabrielle Ras, Ning Xie, Marcel Van Gerven, and Derek Doran. 2022. Explainable deep learning: A field guide for the uninitiated. *Journal of Artificial Intelligence Research*, 73:329–396.
- Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. Zeroscrolls: A zero-shot benchmark for long text understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7977–7989.
- Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. 2006. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence*, pages 1015–1021. Springer.
- Weiwei Sun, Lingyong Yan, Zheng Chen, Shuaiqiang Wang, Haichao Zhu, Pengjie Ren, Zhumin Chen, Dawei Yin, Maarten Rijke, and Zhaochun Ren. 2024. Learning to tokenize for generative retrieval. *Advances in Neural Information Processing Systems*, 36.
- Thibaut Thonet, Jos Rozen, and Laurent Besacier. 2024. Elitr-bench: A meeting assistant benchmark for long-context language models. *arXiv preprint arXiv:2403.20262*.
- Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. [Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Retrieval meets long context large language models. *arXiv preprint arXiv:2310.03025*.
- Fangkai Yang, Pu Zhao, Zezhong Wang, Lu Wang, Jue Zhang, Mohit Garg, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. 2023. Empower large language model to perform better on industrial domain-specific question answering. *arXiv preprint arXiv:2305.11541*.
- Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with bertserini. *arXiv preprint arXiv:1902.01718*.
- Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*.
- Jun Zhang, Shuyang Jiang, Jiangtao Feng, Lin Zheng, and Lingpeng Kong. 2023. Cab: comprehensive attention benchmarking on long sequence modeling. In *International Conference on Machine Learning*, pages 41194–41218. PMLR.
- Zhenyu Zhang, Runjin Chen, Shiwei Liu, Zhewei Yao, Olatunji Ruwase, Beidi Chen, Xiaoxia Wu, and Zhangyang Wang. 2024. Found in the middle: How language models use long contexts better via plug-and-play positional encoding. *arXiv preprint arXiv:2403.04797*.
- Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*.

A Detailed Experimental Setup

A.1 Data Statistics

NQ dataset (Kwiatkowski et al., 2019) has a Creative Commons Share-Alike 3.0 license. This license requires that reusers give credit to the creator. It allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, even for commercial purposes. If others remix, adapt, or build upon the material, they must license the modified material under identical terms.

	Train	Test
Size	79,168	3,610
Question Length	9.17	9.22
Answer Length	2.16	2.22
Answers	1.22	1.79

Table 6: Statistics of NQ dataset. We utilize this dataset to construct contexts and validate, considering semantic relevancy and structural diversity.

A.2 Tools

Pinecone used in Section 3.2 is, an open-source vector database, reduces the cost and time associated with text embedding by storing the embedded text.

A.3 Prompt Template

<p>Provide Long Context Context: {{context}}</p> <p>--</p> <p>Task Instruction Generate an answer (A) to the question (Q) based on the given context. Also, provide the evidence paragraph number (Evidence num) that supports your answer. Please respond according to the format below.</p> <p>Output Format Instruction # Q: {{question}} # A: # Evidence num:</p>
--

Table 7: Prompt template example for the QA and evidence selection task.

A.4 Exact Match Score

EM score checks whether the answer is entirely contained within the generated answer. If the an-

swer is included, the score is 1; otherwise 0.

$$\text{EM}_{\text{short}} = \begin{cases} 1, & \text{if answer} \subseteq \text{generated answer} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

B Qualitative Analysis

<p>Title: Code of Federal Regulations</p> <p>... [7] Under the nondelegation doctrine , federal agencies are authorized by enabling legislation to promulgate regulations (rulemaking) . The process of rulemaking is governed by the Administrative Procedure Act (APA) : generally , the APA requires a process that includes publication of the proposed rules in a notice of proposed rulemaking (NPRM) , a period for comments and participation in the decisionmaking , and adoption and publication of the final rule , via the Federal Register.</p> <p>...[8] The Code of Federal Regulations (CFR) is the codification of the general and permanent rules and regulations (sometimes called administrative law) published in the Federal Register by the executive departments and agencies of the federal government of the United States . The CFR is divided into 50 titles that represent broad areas subject to federal regulation. ...</p>
<p>Question: Where are the rules of federal administrative agencies published?</p>
<p>GT paragraph</p> <p>[8] The Code of Federal Regulations (CFR) is the codification of the general and permanent rules and regulations (sometimes called administrative law) published in the Federal Register by the executive departments and agencies of the federal government of the United States . The CFR is divided into 50 titles that represent broad areas subject to federal regulation.</p>
<p>GPT-3.5 Prediction</p> <p>The rules of federal administrative agencies are published in the Code of Federal Regulations (CFR). The CFR is the codification of the general and permanent rules and regulations published in the Federal Register by the executive departments and agencies of the federal government. It is divided into 50 titles that represent broad areas subject to federal regulation. (Evidence num: 8)</p>
<p>Mixtral Prediction</p> <p>A: The rules of federal administrative agencies are published in the Code of Federal Regulations (CFR). Evidence num: 8 The CFR is the codification of the general and permanent rules and regulations published in the Federal Register by the executive departments and agencies of the federal government of the United States.</p>
<p>Claude3 Prediction</p> <p>A: The rules and regulations of federal administrative agencies are first published in the Federal Register, as stated in the first paragraph: "The rules and regulations are first promulgated or published in the Federal Register." Evidence num: 0</p>

Table 8: Qualitative results on the impact of related information contexts in QA task. Surrounding contexts composed solely of related information positively impact LLMs' answer generation abilities.

<p>Title: Code of Federal Regulations</p> <p>... [unrel] [4] J.T. Hellstrom is introduced in 1999 as a high school friend of Billy Abbott , Raul Guittierez , Mackenzie Browning , Brittany Hodges and Rianna Miner . He has a reputation as a ladies ' man . He dates Rianna , who he treats poorly and cheats on with Brittany . They break up and Rianna moves on with Raul. ... [unrel] [5] The Code of Federal Regulations (CFR) is the codification of the general and permanent rules and regulations (sometimes called administrative law) published in the Federal Register by the executive departments and agencies of the federal government of the United States . The CFR is divided into 50 titles that represent broad areas subject to federal regulation . [unrel] [6] J.T. begins a relationship with Billy 's niece , Colleen Carlton . Colleen is heartbroken when J.T. has a one night stand with Brittany 's mother ...</p>
<p>Question: Where are the rules of federal administrative agencies published?</p>
<p>GT paragraph</p> <p>[5] The Code of Federal Regulations (CFR) is the codification of the general and permanent rules and regulations (sometimes called administrative law) published in the Federal Register by the executive departments and agencies of the federal government of the United States . The CFR is divided into 50 titles that represent broad areas subject to federal regulation.</p>
<p>GPT-3.5 Prediction</p> <p>A: The rules of federal administrative agencies are published in the Code of Federal Regulations (CFR). Evidence num: 5</p>
<p>Mixtral Prediction</p> <p>A: The rules of federal administrative agencies are published in the Code of Federal Regulations (CFR). Evidence num: [5]</p>
<p>Claude3 Prediction</p> <p>A: The rules and regulations of federal administrative agencies are published in the Code of Federal Regulations (CFR). Evidence num: 5</p>

Table 9: Qualitative results on the impact of unrelated information contexts in QA task. Surrounding contexts composed solely of unrelated information negatively impact LLMs' answer generation abilities.

<p>Title: Flora Shaw , Lady Lugard [1] Flora Louise Shaw , DBE (born 19 December 1852 – 25 January 1929) , was a British journalist and writer . She is credited with having coined the name “ Nigeria. ... [2] ... The lengthy articles in a leading daily newspaper reveal a late - Victorian metropolitan imagery of colonial space and time . Shaw projected vast empty spaces awaiting energetic English settlers and economic enterprise ...</p>
<p>Question: Who is the person that gave nigeria name?</p>
<p>GT paragraph [1] Flora Louise Shaw , DBE (born 19 December 1852 – 25 January 1929) , was a British journalist and writer . She is credited with having coined the name Nigeria .</p>
<p>GPT-3.5 Prediction Flora Louise Shaw , DBE, is the person who coined the name "Nigeria".</p>
<p>Mixtral Prediction A: Flora Louise Shaw , DBE Evidence num: 18 Flora Louise Shaw, also known as Flora Shaw, is the person who gave Nigeria its name. She suggested the name "Nigeria" for the British Protectorate on the Niger River in an essay that first appeared in The Times on 8 January 1897. She was in search of a new name and coined "Nigeria", in preference to terms such as "Central Sudan", which were associated with the area by some geographers and travelers.</p>
<p>Claude3 Prediction A: According to the context, Flora Louise Shaw is credited with having coined the name "Nigeria" for the British Protectorate on the Niger River. Evidence num: 18</p>

Table 10: Qualitative results of placing the answer paragraph at the top. Placing the answer paragraph at the top of the context has the most positive effect on the model’s output.

<p>Title: Flora Shaw , Lady Lugard ... [18] In 1905 , Shaw wrote what remains the definitive history of Western Sudan and the modern settlement of Northern Nigeria , A Tropical Dependency : An Outline of the Ancient History of the Western Soudan , With an Account of the Modern Settlement of Northern Nigeria (London : Nisbet , 1905) [19] Flora Louise Shaw , DBE (born 19 December 1852 – 25 January 1929) , was a British journalist and writer . She is credited with having coined the name Nigeria. ...</p>
<p>Question: Who is the person that gave nigeria name?</p>
<p>GT paragraph [19] Flora Louise Shaw , DBE (born 19 December 1852 – 25 January 1929) , was a British journalist and writer . She is credited with having coined the name Nigeria .</p>
<p>GPT-3.5 Prediction Evidence num: 17.</p>
<p>Mixtral Prediction A: Flora Shaw , also known as Flora Shaw Lugard, is credited with having coined the name "Nigeria"</p>
<p>Claude3 Prediction A: According to the context, Flora Shaw is the person who coined the name "Nigeria" for the British Protectorate on the Niger River. Evidence num: 17</p>

Table 11: Qualitative results of placing the answer paragraph at the bottom. Placing the answer paragraph at the bottom of the context has negative effect on the model’s output.

<p>Title: Jingle Bell Rock</p> <p>... [9] [unrel] Wikiversity has steam tables with figures and Matlab code ... [unrel] [10] Jingle Bell Rock is an American popular Christmas song first released by Bobby Helms in 1957 (after it was recorded in October 1957) . It has received frequent airplay in the United States during every Christmas season since then . Jingle Bell Rock was composed by Joseph Carleton Beal (1900 – 1967) and James Ross Boothe (1917 – 1976) . Beal was a Massachusetts - born public relations professional and longtime resident of South Ocean Avenue in Atlantic City , New Jersey , and Boothe was an American writer in the advertising business . [11] [unrel] Steam 's capacity to transfer heat is also used in the home : for cooking vegetables , steam cleaning of fabric ...</p>
<p>Question: Who sings the original jingle bell rock song?</p>
<p>GT paragraph</p> <p>[10] Jingle Bell Rock is an American popular Christmas song first released by Bobby Helms in 1957 (after it was recorded in October 1957) . It has received frequent airplay in the United States during every Christmas season since then . Jingle Bell Rock was composed by Joseph Carleton Beal (1900 – 1967) and James Ross Boothe (1917 – 1976) . Beal was a Massachusetts - born public relations professional and longtime resident of South Ocean Avenue in Atlantic City , New Jersey , and Boothe was an American writer in the advertising business .</p>
<p>GT paragraph index number</p> <p>10</p>
<p>GPT-3.5 Prediction</p> <p>A: Bobby Helms sings the original "Jingle Bell Rock" song. Evidence num: 10</p>
<p>Mixtral Prediction</p> <p>A: The original "Jingle Bell Rock" song was first released by Bobby Helms in 1957. Evidence num: 10</p>
<p>Claude3 Prediction</p> <p>A: According to the context, the original "Jingle Bell Rock" song was first released by Bobby Helms in 1957. Evidence num: 10</p>

Table 12: Qualitative results on the impact of related information contexts in the evidence selection task. When the context consists entirely of unrelated information, the evidence selection ability of LLMs improves.

<p>Title: Jingle Bell Rock</p> <p>... [6] After the song was featured on the soundtrack album to the 1996 film Jingle All the Way , the original Bobby Helms version returned to the Billboard country singles charts in late 1996 and early 1997 , reaching a peak of No. 60 . [7] The song is featured in the opening of Lethal Weapon (1987) . It 's also featured in Mean Girls (2004) . [8] Jingle Bell Rock is an American popular Christmas song first released by Bobby Helms in 1957 (after it was recorded in October 1957) . It has received frequent airplay in the United States during every Christmas season since then . Jingle Bell Rock was composed by Joseph Carleton Beal (1900 – 1967) and James Ross Boothe (1917 – 1976) . Beal was a Massachusetts - born public relations professional and longtime resident of South Ocean Avenue in Atlantic City , New Jersey , and Boothe was an American writer in the advertising business .</p>
<p>Question: Who sings the original jingle bell rock song?</p>
<p>GT paragraph</p> <p>[8] Jingle Bell Rock ” is an American popular Christmas song first released by Bobby Helms in 1957 (after it was recorded in October 1957) . It has received frequent airplay in the United States during every Christmas season since then . “ Jingle Bell Rock ” was composed by Joseph Carleton Beal (1900 – 1967) and James Ross Boothe (1917 – 1976) . Beal was a Massachusetts - born public relations professional and longtime resident of South Ocean Avenue in Atlantic City , New Jersey , and Boothe was an American writer in the advertising business .</p>
<p>GT paragraph index number</p> <p>8</p>
<p>GPT-3.5 Prediction</p> <p>The original version of "Jingle Bell Rock" was sung by Bobby Helms. Evidence num: 0</p>
<p>Mixtral Prediction</p> <p>A: Bobby Helms Evidence num: 1, 3, 4, 5, 8 Bobby Helms is the original singer of "Jingle Bell Rock". This is stated in paragraph 1, where it mentions Helms' version from 1957, and further confirmed in paragraph 3 with the charting details of his version. Paragraphs 4 and 5 provide more recent instances of the song's popularity, and paragraph 8 concludes with the songwriting credits, further solidifying Helms' association with the original "Jingle Bell Rock".</p>
<p>Claude3 Prediction</p> <p>A: The original "Jingle Bell Rock" song was sung by Bobby Helms. Evidence num: [1]</p>

Table 13: Qualitative results of the impact of unrelated information contexts in the evidence selection task. When the context is composed of related information, the evidence selection ability of LLMs decreases.

<p>Title: Taijin kyofusho</p> <p>... [3] [unrel] Fifteen years later , Mia is happily married and has a son and successful career . When she goes on a business trip , Rob meets her at her hotel . He shows her a news article about the dead cyclist , whose wife believes him to be alive and is still looking for him . Rob wants to write an anonymous letter telling the truth , but Mia is afraid the letter will be traced . An argument ensues and she kills Rob . She is further shaken upon seeing a road accident outside , where a self - driving pizza delivery truck has hit a pedestrian . Mia rents a pay per view porn movie to provide herself with an alibi ... [unrel] [4] T Tauri stars (TTS) are a class of variable stars named after their prototype – T Tauri . They are found near molecular clouds and identified by their optical variability and strong chromospheric lines . T Tauri stars are pre-main - sequence stars in the process of contracting to the main sequence along the Hayashi track , a luminosity – temperature relationship obeyed by infant stars of less than 3 solar masses (M) in the pre-main - sequence phase of stellar evolution . It ends when a star of 0.5 M develops a radiative zone , or when a larger star commences nuclear fusion on the main sequence . [5] Analogs of T Tauri stars in the higher mass range (2 – 8 solar masses) – A and B spectral type pre – main - sequence stars , are called Herbig Ae ...</p> <hr/> <p>Question: Where are t-tauri stars found on the hr diagram?</p> <hr/> <p>GT paragraph</p> <p>[4] T Tauri stars (TTS) are a class of variable stars named after their prototype – T Tauri . They are found near molecular clouds and identified by their optical variability and strong chromospheric lines . T Tauri stars are pre-main - sequence stars in the process of contracting to the main sequence along the Hayashi track , a luminosity – temperature relationship obeyed by infant stars of less than 3 solar masses (M) in the pre-main - sequence phase of stellar evolution . It ends when a star of 0.5 M develops a radiative zone , or when a larger star commences nuclear fusion on the main sequence.</p> <hr/> <p>GT paragraph index number</p> <p>4</p> <hr/> <p>GPT-3.5 Prediction</p> <p>Evidence num: 4 T Tauri stars are found on the pre-main sequence phase of stellar evolution, specifically along the Hayashi track. This is a luminosity-temperature relationship followed by infant stars of less than 3 solar masses in their pre-main sequence phase.</p> <hr/> <p>Mixtral Prediction</p> <p>A: T Tauri stars are found in the pre-main sequence stage of stellar evolution, along the Hayashi track on the Hertzsprung-Russell (HR) diagram. This is a luminosity-temperature relationship observed in infant stars of less than 3 solar masses. They are located in the lower left part of the HR diagram, characterized by low temperatures and high luminosities. Evidence num: [4], [5]</p> <hr/> <p>Claude3 Prediction</p> <p>A: T Tauri stars are found near the pre-main sequence region of the Hertzsprung-Russell (HR) diagram. They are in the process of contracting towards the main sequence along the Hayashi track. Evidence num: 4 The context states that "T Tauri stars are pre-main-sequence stars in the process of contracting to the main sequence along the Hayashi track, a luminosity–temperature relationship obeyed by infant stars of less than 3 solar masses (M) in the pre-main-sequence phase of stellar evolution."</p>
--

Table 14: Qualitative results of misalignment in mixed setting. Task alignment discrepancy is most pronounced at 8k length mixed information contents.