

Lost in the Middle, and In-Between: Enhancing Language Models' Ability to Reason Over Long Contexts in Multi-Hop QA

[▽]George Arthur Baker, Ankush Raut, [▽]Sagi Shaier

[†]Lawrence E Hunter, Katharina von der Wense^{▽◊}

[▽]University of Colorado Boulder

[†]University of Chicago, Department of Pediatrics

[◊]Johannes Gutenberg University Mainz

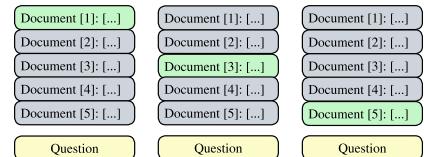
E-mail: {george.baker, sagi.shaier, katharina.kann}@colorado.edu

Abstract

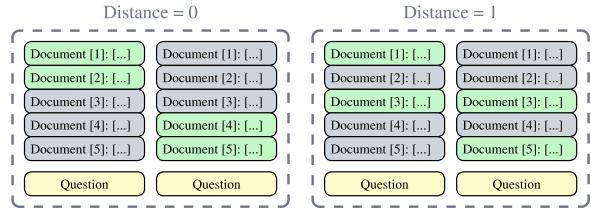
Previous work finds that recent long-context language models fail to make equal use of information in the middle of their inputs, preferring pieces of information located at the tail ends which creates an undue bias in situations where we would like models to be equally capable of using different parts of the input. Thus far, the problem has mainly only been considered in settings with single pieces of critical information, leading us to question what happens when multiple necessary pieces of information are spread out over the inputs. Here, we demonstrate the effects of the "lost in the middle" problem in the multi-hop question answering setting — in which multiple reasoning "hops" over disconnected documents are required — and show that performance degrades not only with respect to the distance of information from the edges of the context, but also between pieces of information. Additionally, we experiment with means of alleviating the problem by reducing superfluous document contents through knowledge graph triple extraction and summarization, and prompting models to reason more thoroughly using chain-of-thought prompting. We make our code and data available at: <https://github.com/Spongeorge/long-context-multihop>

1 Introduction

Recent advancements in attention mechanisms, such as Flash Attention (Dao et al., 2022; Dao, 2023) and Attention with Linear Biases (Press et al., 2022), have ushered in a new generation of language models capable of handling significantly larger context sizes. These developments enable question-answering tasks to be performed over a substantial number of retrieved documents within a single input prompt (as shown in Figure 1). However, despite this remarkable progress, recent studies reveal a critical limitation: long-context models fail to utilize information within their inputs equitably, exhibiting a pronounced bias toward



(a) Previous work mainly explores the "lost in the middle" problem with single-hop questions, which contain just one gold document with the final answer, requiring minimal reasoning.



(b) We explore models' performance with respect to the positions of multiple evidence documents, all of which must be reasoned over to answer a question correctly.

Figure 1: Question Answering setups with documents containing relevant information (green) and distractor documents (gray) placed at different ordinal positions, with both Single-hop (1a) and Multi-hop (1b) questions.

information located at the edges of the context—a problem known as the "lost in the middle" (Liu et al., 2024).

This limitation poses significant challenges for retrieval-augmented generation (RAG) systems, particularly in open-book question-answering scenarios. As the volume of input information increases, so does the likelihood that critical pieces of information necessary for answering a question may be overlooked (Shaier et al., 2024d). This stands in contrast to earlier RAG systems, where documents were processed independently without explicit ordering, resulting in performance that typically improved with the inclusion of more retrieved documents (Izacard and Grave, 2020).

Current efforts to address the "Lost in the Middle" problem primarily focus on approaches like document re-ranking (Peysakhovich and Lerer,

2023; Tang et al., 2023), document length reduction via summarization (Kim et al., 2024), or extending training to include long-context tasks (An et al., 2024). However, these strategies face significant limitations in multi-hop QA settings. In such scenarios, where reasoning involves multiple steps across disconnected documents, the number of possible document order permutations grows combinatorially with the number of reasoning steps. This makes re-ranking and extended training approaches increasingly impractical as they would need to account for an overwhelming number of positional combinations. Similarly, the likelihood of omitting essential information through summarization grows with the number of reasoning steps, jeopardizing the integrity of reasoning chains.

In this work, we investigate the "Lost in the Middle" problem within the context of multi-hop QA. We argue that addressing this issue is critical for advancing the field, given the unique challenges it presents to current mitigation strategies. Specifically, we experiment with chain-of-thought (CoT) prompting (Zhou et al., 2023) and document-size reduction methods to tackle this problem, utilizing recent long-context models such as GPT-3.5-Turbo, MPT-7b-instruct, and Llama-2-7b-longlora. Our key findings include:

1. Performance degradation is not only influenced by the absolute position of information within the context but also by the relative distance between multiple relevant documents.
2. Chain-of-Thought prompting aids in identifying relevant documents but fails to resolve the performance disparity caused by evidence document positions.
3. Existing context reduction methods often produce reasoning chains that are too fragile for effective application in multi-hop QA.

By highlighting these challenges and evaluating potential solutions, this study aims to guide future research in overcoming the "Lost in the Middle" problem and improving long-context model performance in complex multi-hop reasoning tasks.

2 Related Work

2.1 The "Lost in the Middle" Problem

The "Lost in the Middle" problem, first identified by Liu et al. (2024), highlights a significant limitation in long-context LMs. Specifically, when relevant

information is distributed throughout a long context, model performance varies depending on the information's position. Their findings reveal that performance follows a characteristic curve: accuracy is poorest when critical information appears in the middle of the context and improves when the information is near the beginning or end.

In their experiments with the NaturalQuestions-Open dataset (Kwiatkowski et al., 2019), Liu et al. (2024) tested question-answering accuracy by repositioning the document containing the answer among distractor documents. They observed consistent variations in accuracy based on the document's position. Additionally, the authors studied a long-context key-value retrieval task, where models were tasked with retrieving a specific value associated with a key in an extended JSON file. Across both tasks, their results demonstrated that no examined model could process relevant information equally well across all positions.

2.2 Mitigation Strategies for "Lost in the Middle"

To mitigate the "Lost in the Middle" problem, Liu et al. (2024) proposed Query-Aware Contextualization, which involves placing a query both before and after the request. While this approach effectively resolves the issue in the key-value retrieval setting, it has little impact on multi-document question answering, leaving the problem unresolved in that domain.

Other efforts have focused on re-ranking passages before including them in the input prompt. Peysakhovich and Lerer (2023) observed that LLMs assign preferential attention to relevant documents compared to irrelevant ones, even when located at the same position. They proposed sorting documents based on average attention scores while accounting for the typical attention distribution associated with positional biases.

Similarly, Tang et al. (2023) introduced "permutation self-consistency," a method that shuffles document orders and asks the model to rank their relevance, using a cumulative vote to determine the final order. However, these approaches are likely to scale poorly in the multi-hop QA setting. In multi-hop reasoning, later documents in the chain depend on earlier ones, making the number of permutations required for a robust ranking grow combinatorially with the number of reasoning steps.

2.3 Multi-Hop Question Answering

Multi-Hop Question Answering (MHQA) tasks involve reasoning across multiple documents (Yang et al., 2018; Saxena et al., 2020; Mavi et al., 2024), where relevant information is often distributed across the context in disconnected pieces. This often requires models to combine parametric knowledge (Guo et al., 2022; Feng et al., 2023; Shaier et al., 2024a, 2023a; Trivedi et al., 2020; Su et al., 2024; Lee et al., 2021) with complex external context to derive answers. Unlike simpler QA tasks where all relevant information is co-located, the distributed nature of multi-hop reasoning poses significant challenges, often resulting in degraded performance.

The complexity of traversing reasoning chains across multiple sources not only impacts factuality (Guo et al., 2022; Pezeshkpour, 2023; Wang et al., 2023; Shaier et al., 2023b; Wang et al., 2024; Su et al., 2024) but also hinders the model’s ability to consistently utilize all relevant information (Yang et al., 2018; Su et al., 2024; Shaier et al., 2024b,c; Trivedi et al., 2020). As input size and reasoning steps grow, maintaining factual accuracy becomes increasingly difficult, further compounding these challenges.

2.4 Impact of Input Length on Reasoning

Simultaneously to our work, Levy et al. (2024) examined how LLM performance degrades with increasing input length. They found that reasoning performance deteriorates as the number of input tokens grows.

Our study differs from theirs in several key aspects:

1. We focus on performance with respect to document *position* within a fixed-size context, whereas Levy et al. (2024) investigate performance relative to overall input size.
2. We evaluate on three popular multi-hop QA datasets, while Levy et al. (2024) use their custom dataset, FLenQA, which consists exclusively of true/false questions.
3. We study questions requiring up to four reasoning steps, whereas Levy et al. (2024) limit their analysis to two-step comparison questions.

These distinctions emphasize our focus on understanding how positional biases within fixed contexts impact reasoning, particularly in multi-hop QA tasks.

Dataset	Hops	Questions
HotpotQA	2	3703
2WikiMultihopQA ¹	2, 4	6288
MuSiQue-Ans	2, 3, 4	1209

Table 1: Multi-hop datasets we use to evaluate the Lost in the Middle problem. We use the 2nd half of the validation sets due to private test set labels.

texts impact reasoning, particularly in multi-hop QA tasks.

3 Experiments

3.1 Datasets

To evaluate language models on the "Lost in the Middle" problem in the multi-hop setting, we utilize existing Multi-hop Question Answering datasets (Table 1). These datasets allow us to systematically position documents containing relevant information at various locations within the context, interspersed with distractor documents, to analyze how positional biases affect model performance.

Since the official test sets for all three datasets are private and reserved for leaderboard purposes, we create our own splits by dividing the existing validation sets in half. The first half serves as our validation data, while the second half is used as our test set for reporting results.

3.1.1 Models

To investigate the effects of the distance and position of evidence documents within a context on long-context language models, we experiment with a combination of popular open-source and closed-source models. Specifically, we use:

- **MPT-7b-8k-instruct:** An instruction-tuned model trained with ALiBi (Press et al., 2022), which replaces traditional positional embeddings.
- **Llama-2-7b-longlora-8k-ft** (Chen et al., 2023): A fine-tuned version of Llama 2 (Touvron et al., 2023) designed to support long contexts, without instruction tuning.
- **GPT-3.5-turbo-1106:** One of the latest versions of OpenAI’s GPT-3.5 Turbo, offering a context window of 16k tokens and newly reproducible outputs.

¹As 2WikiMultihopQA contains only 10 documents per question, we retrieve an additional 10 distractor documents using a retriever setup as in Liu et al. (2024).

These models, representing a mix of open- and closed-source architectures, allow us to assess the generalizability of our findings across different LLMs.

3.1.2 Metrics

Following Liu et al. (2024), Kandpal et al. (2023), and Mallen et al. (2022), we adopt best-subspan accuracy as our evaluation metric. This metric assigns a score of 1 if the model’s output contains the annotated answer (or any of the alternative answers, in the case of the MuSiQue dataset), and 0 otherwise.

3.1.3 Context Reduction Methods

To investigate the relationship between context size and the "Lost in the Middle" problem, we extend our evaluation by applying two document size reduction methods:

- Knowledge Graph Triple Extraction:** A technique that condenses documents into structured triples, capturing key facts while minimizing extraneous information.
- Document Summarization:** A method that generates concise summaries of documents, preserving their core content while reducing their overall length.

By incorporating these reduction techniques, we aim to explore how modifying context size impacts the manifestation of the "Lost in the Middle" problem.

Knowledge Graph Triple Extraction For extracting knowledge graph triples, we employ an instruction-based approach using the 7-billion-parameter version of LLaMA 2 (Touvron et al., 2023). The model is prompted to extract triples from each individual document in the QA datasets, with the aim of capturing key factual relationships in a structured format.

Summarization To summarize the documents in our datasets, we use BART-large-CNN (Lewis et al., 2019), a pre-trained sequence-to-sequence model fine-tuned on the CNN/Daily Mail news summarization dataset (Hermann et al., 2015). Each document is processed independently, with the maximum generation length capped at 50 tokens to ensure concise summaries.

4 Methodology

For each question in the datasets (and their context-reduced variants), we create multiple prompts by

Dataset	Full	Summ.	KG
HotpotQA	69	29	33
2WikiMultihopQA	45	21	29
MuSiQue-Ans	85	32	35

Table 2: Average document-wise word counts for each dataset and context-reduction method we use.

positioning evidence documents at various locations within a total of 20 documents, following the approach outlined by Liu et al. (2024). Given the high number of potential combinations of evidence positions ($\binom{n}{k} = 190,1140,4845$ possible orderings per prompt in our 2-, 3-, and 4-hop settings, where n represents the number of document positions and k the number of evidence documents), we select a subset of combinations to manage the computational cost of generating long-context prompts. Specifically, we choose 5 combinations where the gold documents are placed adjacent to one another, as well as 4 combinations with distractor documents interspersed between the gold documents for 2- and 4-hop questions, and 3 combinations for 3-hop questions. The distractor documents are retained in their original order, based on their relevance determined through the retrieval process.

The prompts are processed by the models with a temperature setting of 0, and the generation is limited to a maximum of 256 tokens.

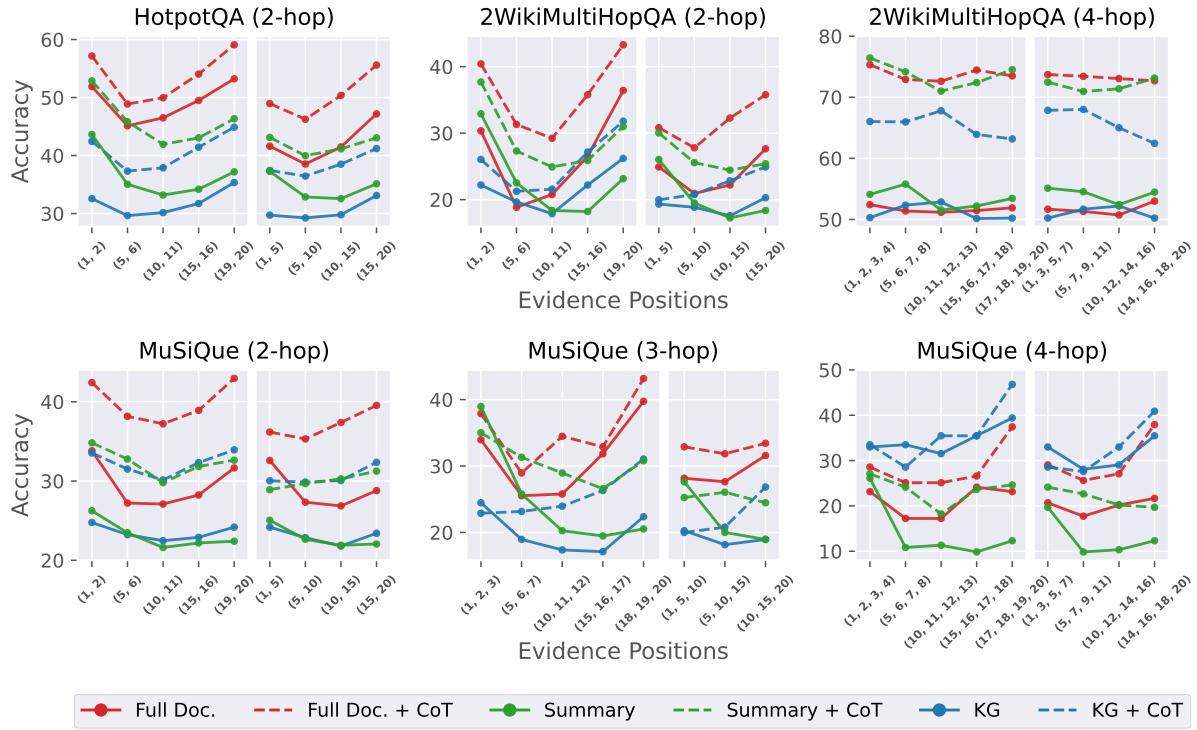
5 Results

The results for instruction-tuned models are presented in Figure 2, while those for non-instruction-tuned models are shown in Figure 3. For a comprehensive view of all results, including instances where models performed suboptimally, please refer to Appendix A. We focus here on the most notable findings.

Specifically, Figures 2 and 3 illustrate the performance variations across the three models, with respect to each dataset, as we manipulate the placement of the relevant documents (e.g., positions 1 and 2, 5 and 6, etc.). Additionally, the figures compare model performance when using the full document, a summarized version, or knowledge graph triples extracted from the document. For each condition, we experiment with and without CoT prompting to assess its impact on model performance.

Lastly, Figure 4 provides a detailed analysis of how the distance between relevant documents im-

mpt-7b-8k-instruct QA Accuracy



gpt-3.5-turbo-1106 QA Accuracy

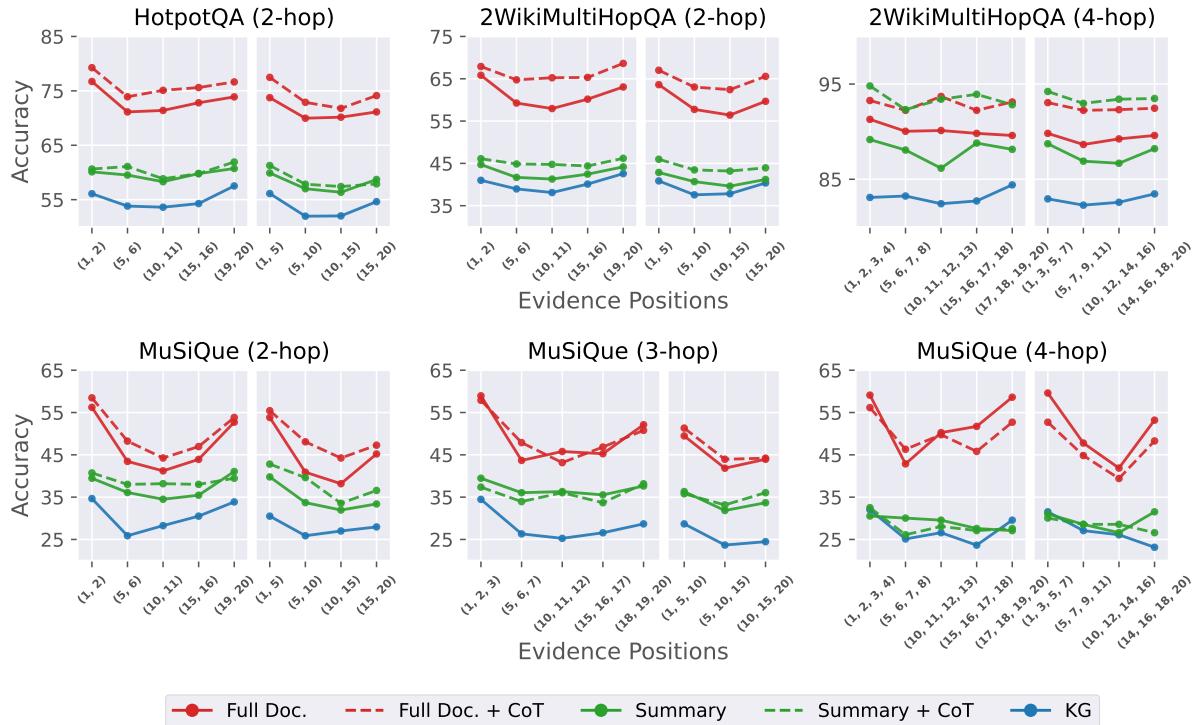


Figure 2: The performance impacts of varying the positions of relevant documents within instruction-tuned models' inputs, with context reduction techniques and Chain-of-Thought prompting. All positions are out of 20 total documents. KG + CoT results for gpt-3.5-turbo are omitted to Appendix A to highlight other results.

Llama-2-7b-longlora-8k-ft QA Accuracy

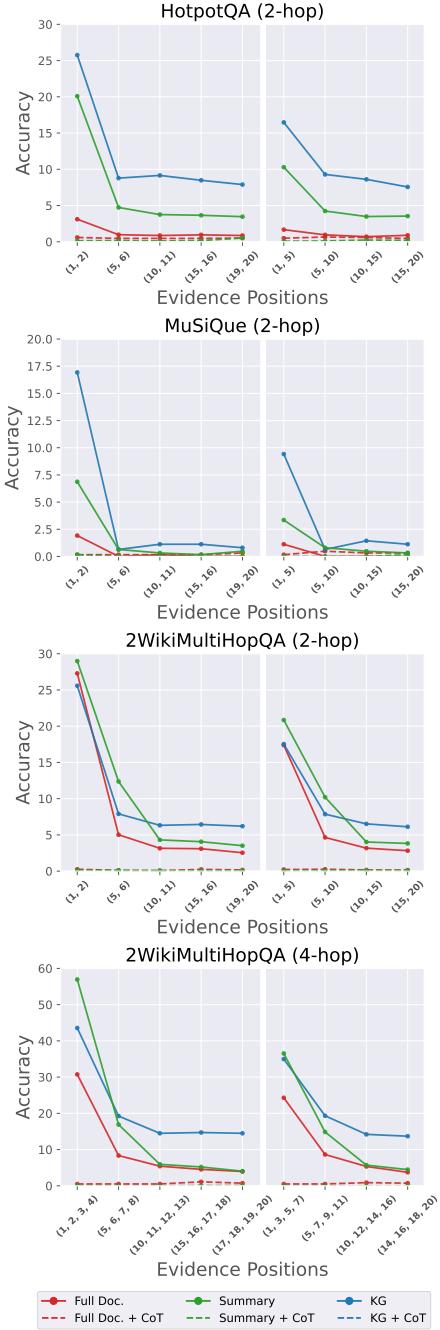


Figure 3: Experimental results for Llama-2-7b-longlora-8k-ft. Results for MuSiQue 3- and 4-hop splits are relegated to Appendix A due to exceedingly poor performance.

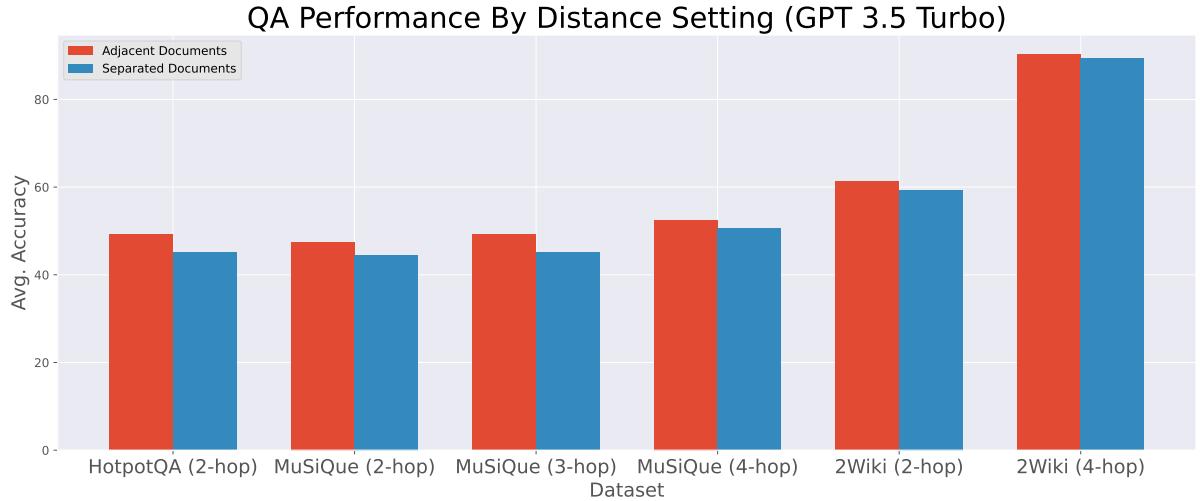
pacts model performance across different datasets. Specifically, it examines the performance variations when relevant documents are placed adjacently at different positions, compared to scenarios where they are separated by distractor documents. This analysis highlights the sensitivity of the models to the spatial arrangement of evidence and underscores the challenges posed by non-adjacent evidence placement in the context of multi-hop reasoning tasks.

6 Analysis

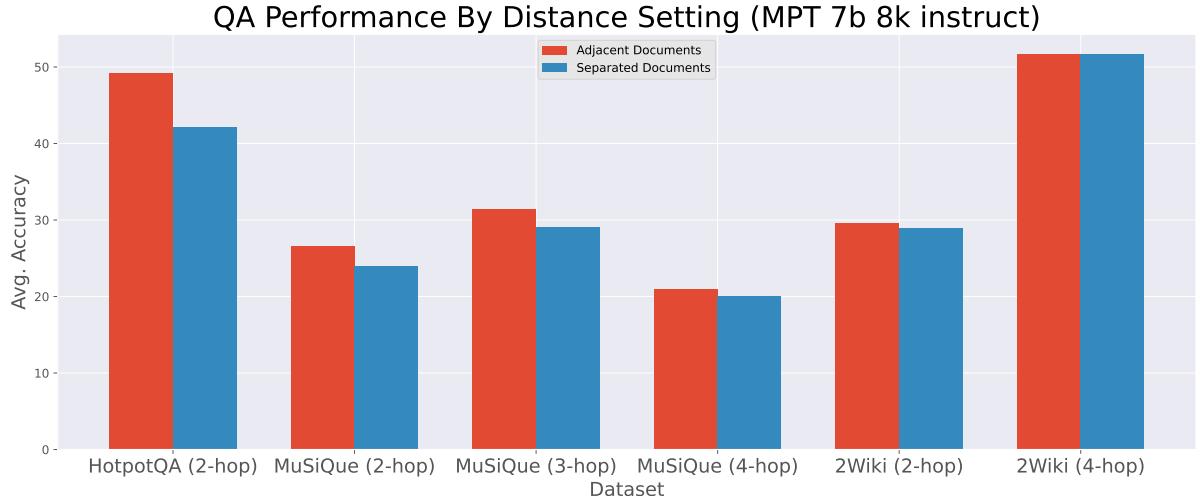
Proximity of relevant documents significantly affects performance. Figure 4 highlights a clear trend: models perform better when relevant documents are adjacent compared to when they are separated by distractor documents. This suggests that the spatial proximity of evidence is crucial for models to effectively retrieve and integrate information. When documents are adjacent, connections between them are more easily captured, potentially due to the model’s attention mechanisms. In contrast, when distractors separate the relevant documents, the models struggle to retrieve and synthesize the necessary evidence, resulting in a drop in performance.

Chain-of-Thought prompting yields mixed results. For instruction-tuned models such as MPT and GPT-3.5, CoT prompting with few-shot exemplars markedly improves performance compared to zero-shot settings in most scenarios. This improvement likely stems from the explicit reasoning steps provided by CoT, which help these models better structure their responses and navigate complex multi-hop reasoning tasks. However, for the non-instruction-tuned Llama 2 longlora model, CoT prompting leads to a sharp decline in performance. This discrepancy may stem from the model’s inherent biases: it demonstrates a pronounced primacy bias, with little to no recency bias, leading to an over-reliance on the few-shot exemplars and improper integration of the actual task-specific context. This suggests that non-instruction-tuned models may require careful tuning or additional training to fully leverage CoT-style reasoning.

Context reduction mitigates position biases but sacrifices accuracy. Figure 2 reveals that reducing context—whether through summarization or knowledge graph triple extraction—dampens the impact of the “lost in the middle” problem. Specifically, evidence located in the middle of the input



(a) Average question-answering accuracy for GPT models with full document prompts by distance setting. Performance is generally higher when evidence documents are adjacent compared to when they are separated by distractor documents.



(b) Average question-answering accuracy for MPT models with full document prompts by distance setting. Similar to GPT, performance decreases as the separation between evidence documents increases.

Figure 4: Average question-answering accuracy for full document prompts by distance setting for GPT and MPT models. Performance with adjacent evidence documents is generally higher than when evidence documents are separated by distractor documents.

achieves performance levels closer to those of edge-positioned evidence when using reduced contexts. This flattening of the performance curve suggests that context reduction alleviates the models' positional biases. However, this improvement comes at a cost: overall accuracy declines, particularly in instruction-tuned models like MPT and GPT-3.5. This drop is likely due to information loss during the context reduction process. Interestingly, the non-instruction-tuned Llama 2 longlora model benefits substantially from context reduction, suggesting that these methods can serve as a useful preprocessing step for less robust models.

7 Future Work

Building on the findings of this study, there are several promising directions for future research aimed at addressing the "Lost in the Middle" problem in multi-hop reasoning tasks.

First, while our work highlights the impact of document positioning and adjacency, a more exhaustive evaluation of evidence combinations is needed. Computational constraints limited our analysis to a subset of possible configurations, leaving unexplored how more complex or nuanced arrangements affect model performance. Advances in more efficient evaluation methodologies or sampling strategies could enable a broader and deeper exploration of this space.

Second, the limitations of current context-reduction techniques, such as summarization and knowledge graph triple extraction, underscore the need for improved preprocessing methods. Future research could develop tailored strategies that preserve key reasoning paths while minimizing extraneous information. Techniques leveraging recent advancements in unsupervised learning, retrieval-augmented generation, or specialized fine-tuning could prove effective in mitigating information loss.

Third, the disparity in performance improvements between instruction-tuned and non-instruction-tuned models suggests an opportunity to better align model architectures with task-specific reasoning demands. Investigating the potential of advanced prompting techniques, such as dynamic CoT, could enhance model performance, particularly for non-instruction-tuned models like Llama 2 longlora.

Additionally, exploring the robustness of newer and larger models to the "Lost in the Middle" prob-

lem remains an open question. Recent advancements in large-scale pretraining and reasoning capabilities may yield models better equipped to handle dispersed evidence, and future evaluations should incorporate these state-of-the-art systems to identify potential improvements.

Finally, incorporating external memory mechanisms or augmenting model architectures to dynamically prioritize and retrieve relevant evidence could offer a path forward. Such modifications may allow models to better handle long-context reasoning challenges, reducing sensitivity to document positioning and improving overall robustness in multi-hop settings.

By addressing these directions, future research can move closer to resolving the challenges posed by dispersed evidence in multi-hop question answering and enhancing the capabilities of long-context language models in reasoning-intensive tasks.

8 Conclusion

In this paper, we presented a study of the "Lost in the Middle" problem in the context of Multi-hop Question Answering, where models must integrate information from multiple documents to generate correct answers. Using three widely-used multi-hop QA datasets (HotpotQA, 2WikiMultihopQA, and MuSiQue-Ans), we analyzed the performance of recent large language models as a function of the positions of evidence documents within a context interspersed with distractors. Our findings reveal that model performance is not only influenced by the absolute positions of evidence documents but also by their relative positioning, highlighting a previously underexplored dimension of this problem.

We also explored context-reduction techniques such as summarization and knowledge graph triple extraction as potential solutions. However, our results indicate that these out-of-the-box approaches are insufficient to fully mitigate the issue, as they often lead to a trade-off between reducing positional bias and retaining critical information.

Overall, our work underscores the complexity of the "Lost in the Middle" problem in multi-hop settings, extending beyond the single-hop scenarios that current mitigation strategies typically address. This study opens up new challenges for the design of LLMs and methods aimed at improving reasoning across long and complex contexts.

Limitations

Our evaluation of the "Lost in the Middle" problem in multi-hop settings is subject to several limitations, primarily due to computational and time constraints. First, we assessed a carefully curated subset of possible evidence location combinations rather than exhaustively evaluating all permutations. The factorial growth in combinations, combined with the high token counts of our prompts, made a comprehensive analysis computationally prohibitive. While our approach highlights meaningful patterns, it may not capture the full extent of potential variations in performance.

Second, although the models evaluated were all state-of-the-art at the time of this study and released within the past year, newer and potentially more powerful models have since become available. These more recent models, with greater reasoning capabilities and larger parameter counts, may exhibit improved robustness against the "Lost in the Middle" problem. Investigating the behavior of these models remains an important direction for future work.

Lastly, our experiments primarily focus on out-of-the-box summarization and knowledge graph extraction techniques for context reduction. More advanced, model-specific fine-tuning or optimized preprocessing strategies may yield better results, but these were beyond the scope of our current study. Future research could explore such targeted interventions to further address the challenges identified here.

References

- Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, and Jian-Guang Lou. 2024. Make your llm fully utilize the context. *arXiv preprint arXiv:2404.16811*.
- Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2023. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*.
- Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.
- Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359.
- Shangbin Feng, Vidhisha Balachandran, Yuyang Bai, and Yulia Tsvetkov. 2023. FactKB: Generalizable factuality evaluation using language models enhanced with factual knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 933–952, Singapore. Association for Computational Linguistics.
- Wangzhen Guo, Qinkang Gong, and Hanjiang Lai. 2022. Counterfactual multihop qa: A cause-effect approach for reducing disconnected reasoning.
- Karl Moritz Hermann, Tomas Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.
- Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *CoRR*, abs/2007.01282.
- Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.
- Jaehyun Kim, Jaehyun Nam, Sangwoo Mo, Jongjin Park, Sang-Woo Lee, Minjoon Seo, Jung-Woo Ha, and Jinwoo Shin. 2024. Sure: Improving open-domain question answering of LLMs via summarized retrieval. In *The Twelfth International Conference on Learning Representations*.
- 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.
- Kyungjae Lee, Seung won Hwang, Sang eun Han, and Dohyeon Lee. 2021. Robustifying multi-hop qa through pseudo-evidentiality training.
- 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*.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.
- 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.

- Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*.
- Vaibhav Mavi, Anubhav Jangra, and Adam Jatowt. 2024. Multi-hop question answering.
- Alexander Peysakhovich and Adam Lerer. 2023. Attention sorting combats recency bias in long context language models.
- Pouya Pezeshkpour. 2023. Measuring and modifying factual knowledge in large language models.
- Ofir Press, Noah Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*.
- Apoory Saxena, Adityay Tripathi, and Partha Talukdar. 2020. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4498–4507, Online. Association for Computational Linguistics.
- Sagi Shaier, Kevin Bennett, Lawrence Hunter, and Katharina Kann. 2023a. Emerging challenges in personalized medicine: Assessing demographic effects on biomedical question answering systems. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 540–550, Nusa Dua, Bali. Association for Computational Linguistics.
- Sagi Shaier, Kevin Bennett, Lawrence Hunter, and Katharina von der Wense. 2024a. Comparing template-based and template-free language model probing. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 766–776, St. Julian’s, Malta. Association for Computational Linguistics.
- Sagi Shaier, Lawrence Hunter, and Katharina Kann. 2023b. Who are all the stochastic parrots imitating? they should tell us! In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 113–120, Nusa Dua, Bali. Association for Computational Linguistics.
- Sagi Shaier, Lawrence Hunter, and Katharina Wense. 2024b. Desiderata for the context use of question answering systems. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 777–792, St. Julian’s, Malta. Association for Computational Linguistics.
- Sagi Shaier, Lawrence Hunter, and Katharina Wense. 2024c. It is not about what you say, it is about how you say it: A surprisingly simple approach for improving reading comprehension. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8292–8305, Bangkok, Thailand. Association for Computational Linguistics.
- Sagi Shaier, Ari Kobren, and Philip V. Ogren. 2024d. Adaptive question answering: Enhancing language model proficiency for addressing knowledge conflicts with source citations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17226–17239, Miami, Florida, USA. Association for Computational Linguistics.
- Xin Su, Tiep Le, Steven Bethard, and Phillip Howard. 2024. Semi-structured chain-of-thought: Integrating multiple sources of knowledge for improved language model reasoning.
- Raphael Tang, Xinyu Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Ture. 2023. Found in the middle: Permutation self-consistency improves listwise ranking in large language models.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2020. Is multihop QA in DiRe condition? measuring and reducing disconnected reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8846–8863, Online. Association for Computational Linguistics.
- Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity.
- Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Nenkov Georgiev, Rocktim Jyoti Das, and Preslav Nakov. 2024. Factualty of large language models: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19519–19529, Miami, Florida, USA. Association for Computational Linguistics.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages

2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei,
Nathan Scales, Xuezhi Wang, Dale Schuurmans,
Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi.
2023. [Least-to-most prompting enables complex reasoning in large language models](#).

A Full Results

Dataset (Closed-book Score)	Prompt	1,2	5,6	10,11	15,16	19,20	1,5	5,10	10,15	15,20
Hotpot (40.64%)	standard	76.72%	71.13%	71.40%	72.81%	73.89%	73.75%	69.97%	70.16%	71.13%
	standard + CoT	79.26%	73.91%	75.10%	75.61%	76.64%	77.48%	72.89%	71.78%	74.13%
	kg	56.09%	53.82%	53.61%	54.28%	57.52%	56.12%	51.96%	52.01%	54.63%
	kg + CoT	34.94%	28.03%	29.57%	29.92%	31.14%	31.60%	28.25%	28.54%	29.84%
	summaries	60.11%	59.52%	58.30%	59.79%	60.73%	59.90%	57.01%	56.36%	58.71%
	summaries + CoT	60.63%	61.09%	58.84%	59.79%	61.92%	61.27%	57.84%	57.41%	57.90%
MuSiQue (14.39%)		1,2	5,6	10,11	15,16	19,20	1,5	5,10	10,15	15,20
2-hop	standard	56.23%	43.45%	41.21%	43.93%	52.72%	53.83%	40.89%	38.18%	45.21%
	standard + CoT	58.47%	48.24%	44.25%	46.96%	53.83%	55.43%	48.08%	44.25%	47.28%
	kg	34.66%	25.88%	28.27%	30.51%	33.87%	30.51%	25.88%	27.00%	27.96%
	kg + CoT	14.06%	7.67%	9.90%	9.11%	10.86%	12.14%	8.63%	9.11%	9.90%
	summaries	39.46%	36.10%	34.50%	35.46%	41.05%	39.78%	33.71%	31.95%	33.39%
	summaries + CoT	40.73%	38.02%	38.18%	38.02%	39.46%	42.81%	39.62%	33.55%	36.58%
		1, 2, 3	5, 6, 7	10, 11, 12	15, 16, 17	18, 19, 20	1, 5, 10	5, 10, 15	10, 15, 20	
3-hop	standard	58.95%	43.68%	45.79%	45.26%	52.11%	49.47%	41.84%	43.95%	-
	standard + CoT	57.89%	47.89%	43.16%	46.84%	50.79%	51.32%	43.95%	44.21%	-
	kg	34.47%	26.32%	25.26%	26.58%	28.68%	28.68%	23.68%	24.47%	-
	kg + CoT	5.79%	3.68%	5.79%	5.79%	5.79%	5.53%	4.47%	-	-
	summaries	39.47%	36.05%	36.32%	35.53%	37.63%	36.32%	31.84%	33.68%	-
	summaries + CoT	37.37%	33.95%	36.05%	33.68%	38.16%	35.79%	33.16%	36.05%	-
		1,2,3,4	5,6,7,8	10,11,12,13	15,16,17,18	17,18,19,20	1,3,5,7	5,7,9,11	10,12,14,16	14,16,18,20
4-hop	standard	59.11%	42.86%	50.25%	51.72%	58.62%	59.61%	47.78%	41.87%	53.20%
	standard + CoT	56.16%	46.31%	49.75%	45.81%	52.71%	52.71%	44.83%	39.41%	48.28%
	kg	32.02%	25.12%	26.60%	23.65%	29.56%	31.53%	27.09%	26.11%	23.15%
	kg + CoT	3.94%	3.45%	5.91%	5.91%	4.93%	5.91%	3.94%	5.91%	-
	summaries	30.54%	30.05%	29.56%	27.59%	27.09%	31.03%	28.57%	26.60%	31.53%
	summaries + CoT	32.51%	26.11%	28.08%	27.09%	27.59%	30.05%	28.57%	28.57%	26.60%
2Wiki (44.99%)		1,2	5,6	10,11	15,16	19,20	1,5	5,10	10,15	15,20
2-hop	standard	65.83%	59.26%	57.96%	60.18%	63.06%	63.61%	57.76%	56.42%	59.67%
	standard + CoT	67.90%	64.75%	65.24%	65.32%	68.63%	67.01%	63.04%	62.43%	65.56%
	kg	41.00%	38.97%	38.11%	40.11%	42.58%	40.86%	37.57%	37.83%	40.37%
	kg + CoT	25.99%	24.08%	24.32%	24.46%	26.11%	25.05%	24.38%	23.77%	24.42%
	summaries	44.72%	41.69%	41.30%	42.46%	44.17%	42.85%	40.69%	39.62%	41.20%
	summaries + CoT	46.10%	44.86%	44.76%	44.39%	46.20%	45.98%	43.46%	43.19%	43.95%
		1,2,3,4	5,6,7,8	10,11,12,13	15,16,17,18	17,18,19,20	1,3,5,7	5,7,9,11	10,12,14,16	14,16,18,20
4-hop	standard	91.29%	90.04%	90.12%	89.82%	89.60%	89.82%	88.65%	89.24%	89.60%
	standard + CoT	93.27%	92.24%	93.70%	92.24%	93.12%	93.05%	92.24%	92.31%	92.46%
	kg	83.09%	83.24%	82.43%	82.72%	84.41%	82.94%	82.28%	82.58%	83.46%
	kg + CoT	65.15%	67.20%	67.94%	64.28%	64.13%	66.62%	65.96%	65.81%	63.47%
	summaries	89.17%	88.07%	86.16%	88.80%	88.14%	88.73%	86.90%	86.68%	88.21%
	summaries + CoT	94.80%	92.31%	93.41%	93.92%	92.83%	94.22%	92.97%	93.41%	93.48%

Table 3: Full experimental results of gpt-3.5-turbo-1106. Percentages next to dataset names are the closed-book scores for the full set.

Dataset (Closed-book Score)	Prompt	1,2	5,6	10,11	15,16	19,20	1,5	5,10	10,15	15,20
Hotpot (14.88%)	standard	51.90%	45.13%	46.50%	49.50%	53.23%	41.61%	38.54%	41.48%	47.18%
	standard + CoT	57.17%	48.88%	49.96%	54.04%	59.09%	48.96%	46.26%	50.36%	55.60%
	kg	32.57%	29.65%	30.16%	31.73%	35.35%	29.73%	29.25%	29.79%	33.11%
	kg + CoT	42.45%	37.32%	37.89%	41.43%	44.88%	37.43%	36.46%	38.51%	41.21%
	summaries	43.64%	35.05%	33.19%	34.19%	37.19%	32.77%	32.87%	32.57%	35.13%
	summaries + CoT	52.88%	45.83%	41.94%	43.05%	46.34%	43.13%	39.97%	41.16%	43.05%
Musique (3.64%)		1,2	5,6	10,11	15,16	19,20	1,5	5,10	10,15	15,20
2-hop	standard	30.35%	18.85%	20.77%	26.52%	36.42%	24.92%	20.93%	22.20%	27.64%
	standard + CoT	40.42%	31.31%	29.23%	35.78%	43.29%	30.83%	27.80%	32.27%	35.78%
	kg	22.20%	19.65%	17.89%	22.20%	26.20%	19.33%	18.85%	17.57%	20.29%
	kg + CoT	26.04%	21.25%	21.57%	27.16%	31.79%	19.97%	20.77%	22.84%	24.92%
	summaries	32.91%	22.52%	18.37%	18.21%	23.16%	26.04%	19.49%	17.25%	18.37%
	summaries + CoT	37.70%	27.32%	24.92%	25.88%	30.99%	30.03%	25.56%	24.44%	25.40%
		1,2,3	5,6,7	10,11,12	15,16,17	18,19,20	1,5,10	5,10,15	10,15,20	
3-hop	standard	33.95%	25.53%	25.79%	31.84%	39.74%	28.16%	27.63%	31.58%	-
	standard + CoT	37.89%	28.95%	34.47%	32.89%	43.16%	32.89%	31.84%	33.42%	-
	kg	24.47%	18.95%	17.37%	17.11%	22.37%	20.26%	18.16%	18.95%	-
	kg + CoT	22.89%	23.16%	23.95%	26.32%	31.05%	20.00%	20.79%	26.84%	-
	summaries	38.95%	25.79%	20.26%	19.47%	20.53%	27.63%	20.00%	18.95%	-
	summaries + CoT	35.00%	31.32%	28.95%	26.58%	30.79%	25.26%	26.05%	24.47%	-
		1,2,3,4	5,6,7,8	10,11,12,13	15,16,17,18	17,18,19,20	1,3,5,7	5,7,9,11	10,12,14,16	14,16,18,20
4-hop	standard	23.15%	17.24%	17.24%	24.14%	23.15%	20.69%	17.73%	20.20%	21.67%
	standard + CoT	28.57%	25.12%	25.12%	26.60%	37.44%	29.06%	25.62%	27.09%	37.93%
	kg	33.00%	33.50%	31.53%	35.47%	39.41%	33.00%	30.08%	29.06%	35.47%
	kg + CoT	33.50%	28.57%	35.47%	35.47%	46.80%	28.57%	27.59%	33.00%	40.89%
	summaries	26.11%	10.84%	11.33%	9.85%	12.32%	19.70%	9.85%	10.34%	12.32%
	summaries + CoT	27.09%	24.14%	18.23%	23.65%	24.63%	24.14%	22.66%	20.20%	19.70%
2wiki (20.13%)		1,2	5,6	10,11	15,16	19,20	1,5	5,10	10,15	15,20
2-hop	standard	33.81%	27.22%	27.10%	28.24%	31.65%	32.59%	27.33%	26.86%	28.79%
	standard + CoT	42.42%	38.16%	37.22%	38.91%	42.95%	36.18%	35.31%	37.38%	39.54%
	kg	24.77%	23.24%	22.47%	22.90%	24.18%	24.18%	22.86%	21.80%	23.41%
	kg + CoT	33.54%	31.51%	30.13%	32.32%	33.93%	30.03%	29.87%	30.07%	32.36%
	summaries	26.25%	23.47%	21.62%	22.19%	22.41%	25.05%	22.65%	21.90%	22.06%
	summaries + CoT	34.82%	32.77%	29.80%	31.84%	32.63%	28.91%	29.68%	30.27%	31.25%
		1,2,3,4	5,6,7,8	10,11,12,13	15,16,17,18	17,18,19,20	1,3,5,7	5,7,9,11	10,12,14,16	14,16,18,20
4-hop	standard	52.42%	51.39%	51.17%	51.46%	51.90%	51.68%	51.32%	50.73%	53.00%
	standard + CoT	75.33%	72.91%	72.62%	74.45%	73.50%	73.72%	73.43%	73.06%	72.69%
	kg	50.29%	52.34%	52.86%	50.15%	50.22%	50.22%	51.68%	52.20%	50.22%
	kg + CoT	66.03%	65.96%	67.79%	63.91%	63.18%	67.86%	68.01%	65.01%	62.45%
	summaries	54.10%	55.78%	51.54%	52.20%	53.44%	55.12%	54.54%	52.42%	54.47%
	summaries + CoT	76.43%	74.16%	71.01%	72.40%	74.52%	72.47%	70.94%	71.38%	73.13%

Table 4: Full experimental results of mpt-7b-8k-instruct. Percentages next to dataset names are the closed-book scores for the full set.

Dataset (Closed-book Score)	Prompt	1,2	5,6	10,11	15,16	19,20	1,5	5,10	10,15	15,20
Hotpot (12.21%)	standard	3.11%	0.97%	0.86%	0.95%	0.86%	1.67%	0.95%	0.70%	0.89%
	standard + CoT	0.59%	0.46%	0.46%	0.46%	0.51%	0.49%	0.65%	0.59%	0.46%
	kg	25.76%	8.78%	9.15%	8.48%	7.89%	16.47%	9.29%	8.61%	7.56%
	kg + CoT	1.16%	0.97%	1.03%	1.05%	0.81%	0.89%	1.08%	0.89%	1.08%
	summaries	20.09%	4.73%	3.75%	3.65%	3.46%	10.29%	4.24%	3.48%	3.54%
	summaries + CoT	0.14%	0.16%	0.11%	0.16%	0.51%	0.08%	0.11%	0.24%	0.19%
Musique (0.74%)		1,2	5,6	10,11	15,16	19,20	1,5	5,10	10,15	15,20
2-hop	standard	1.92%	0.00%	0.16%	0.00%	0.00%	1.12%	0.00%	0.00%	0.00%
	standard + CoT	0.16%	0.16%	0.16%	0.16%	0.32%	0.16%	0.48%	0.32%	0.32%
	kg	16.93%	0.64%	1.12%	1.12%	0.80%	9.42%	0.64%	1.44%	1.12%
	kg + CoT	0.16%	0.00%	0.16%	0.16%	0.16%	0.16%	0.16%	0.16%	0.16%
	summaries	6.87%	0.64%	0.32%	0.16%	0.48%	3.35%	0.80%	0.48%	0.32%
	summaries + CoT	0.16%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.16%
		1,2,3	5,6,7	10,11,12	15,16,17	18,19,20	1,5,10	5,10,15	10,15,20	
3-hop	standard	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	-
	standard + CoT	0.00%	0.00%	0.26%	0.00%	0.00%	0.00%	0.26%	0.26%	-
	kg	0.00%	0.00%	0.00%	0.26%	0.26%	0.00%	0.00%	0.26%	-
	kg + CoT	0.26%	0.26%	0.26%	0.26%	0.00%	0.26%	0.26%	0.00%	-
	summaries	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	-
	summaries + CoT	0.00%	0.00%	0.00%	0.00%	0.26%	0.00%	0.00%	0.00%	-
		1,2,3,4	5,6,7,8	10,11,12,13	15,16,17,18	17,18,19,20	1,3,5,7	5,7,9,11	10,12,14,16	14,16,18,20
4-hop	standard	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	standard + CoT	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	kg	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	kg + CoT	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	summaries	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	summaries + CoT	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
2wiki (31.65%)		1,2	5,6	10,11	15,16	19,20	1,5	5,10	10,15	15,20
2-hop	standard	27.29%	5.02%	3.15%	3.09%	2.54%	17.39%	4.65%	3.17%	2.82%
	standard + CoT	0.26%	0.10%	0.06%	0.24%	0.16%	0.22%	0.26%	0.16%	0.14%
	kg	25.58%	7.90%	6.32%	6.44%	6.20%	17.53%	7.86%	6.52%	6.12%
	kg + CoT	0.81%	0.77%	0.89%	1.00%	1.02%	1.00%	1.06%	1.02%	1.02%
	summaries	28.99%	12.37%	4.31%	4.06%	3.51%	20.85%	10.20%	4.02%	3.82%
	summaries + CoT	0.12%	0.10%	0.08%	0.06%	0.06%	0.06%	0.10%	0.10%	0.08%
		1,2,3,4	5,6,7,8	10,11,12,13	15,16,17,18	17,18,19,20	1,3,5,7	5,7,9,11	10,12,14,16	14,16,18,20
4-hop	standard	30.75%	8.35%	5.42%	4.54%	3.95%	24.30%	8.64%	5.34%	3.73%
	standard + CoT	0.51%	0.51%	0.51%	1.10%	0.73%	0.51%	0.51%	0.88%	0.73%
	kg	43.56%	19.25%	14.49%	14.71%	14.49%	34.99%	19.33%	14.20%	13.69%
	kg + CoT	2.56%	3.15%	2.12%	2.34%	2.34%	1.98%	2.64%	2.27%	2.42%
	summaries	56.95%	16.91%	5.93%	5.20%	4.03%	36.53%	14.86%	5.71%	4.47%
	summaries + CoT	0.12%	0.00%	0.00%	0.07%	0.07%	0.06%	0.00%	0.00%	0.00%

Table 5: Full experimental results of llama-2-7b-longlora-8k-ft. Percentages next to dataset names are the closed-book scores for the full set.