Do RAG Systems Really Suffer From Positional Bias?

Florin Cuconasu[1,2*†], **Simone Filice**[2]
Guy Horowitz[2], **Yoelle Maarek**[2], **Fabrizio Silvestri**[1]

[1]Sapienza University of Rome, [2]Technology Innovation Institute

Abstract

Retrieval Augmented Generation enhances LLM accuracy by adding passages retrieved from an external corpus to the LLM prompt. This paper investigates how positional bias–the tendency of LLMs to weight information differently based on its position in the prompt–affects not only the LLM's capability to capitalize on relevant passages, but also its susceptibility to distracting passages. Through extensive experiments on three benchmarks, we show how state-of-the-art retrieval pipelines, while attempting to retrieve relevant passages, systematically bring highly distracting ones to the top ranks, with over 60% of queries containing at least one highly distracting passage among the top-10 retrieved passages. As a result, the impact of the LLM positional bias, which in controlled settings is often reported as very prominent by related works, is actually marginal in real scenarios since both relevant and distracting passages are, in turn, penalized. Indeed, our findings reveal that sophisticated strategies that attempt to rearrange the passages based on LLM positional preferences do not perform better than random shuffling.

1 Introduction

Retrieval Augmented Generation (RAG) improves the factual accuracy of LLMs on knowledge-intensive tasks by including in the prompt passages retrieved from an external corpus (Chen et al., 2017; Petroni et al., 2021b; Fan et al., 2024). Because any real retriever is imperfect, RAG systems feed the LLM *several* top-ranked passages, not just the single best one. That practice raises recall but also inserts *distracting* passages: text that looks relevant yet lacks the appropriate answer. Recent work shows these distractors can sharply degrade the LLM answer accuracy (Cuconasu et al., 2024; Jin et al., 2025; Yoran et al., 2024).

A second, orthogonal weakness of LLMs is *positional bias*: moving the same evidence to a different location in the context can change the answer and largely impact its accuracy. Liu et al. (2024) term this the *lost-in-the-middle* effect, to refer to the tendency of LLMs to focus on text appearing in the beginning or end of their prompt. Prior analyses (Liu et al., 2024; Hutter et al., 2025; He et al., 2024), however, study the problem in a controlled setting, typically rotating the position of a sole relevant passage in a prompt otherwise containing only irrelevant passages. This artificial configuration not only amplifies the impact of the positional bias but also ignores how the positional bias influences the vulnerability of the LLMs to distracting passages, which instead is central in our work.

Using the "distracting effect" metric of Amiraz et al. (2025), we show that answer accuracy depends on the positions of *both* relevant and distracting passages. Then, we empirically show that current state-of-the-art retrieval pipelines, while attempting to retrieve relevant passages, also bring highly distracting passages to the top ranks, and the more advanced the retrieval pipeline is, the more distracting the passages are. This simultaneous presence of relevant and highly distracting passages near the top of the retrieval ranking drastically reduces the impact of the positional bias, since it penalizes, in turn, both passage types.

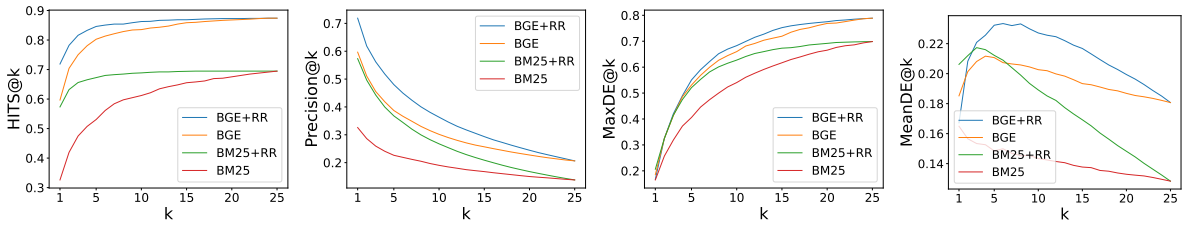
Following these findings, we empirically demonstrate that strategies to rearrange the passages in the prompt based on the LLM-preferred positions are not more effective than a random passage ordering.

2 Related work

Effect of Irrelevant Content. Recent work explores the detrimental effect of irrelevant content in the LLM prompt. In the RAG setting, a passage is considered irrelevant if it does not provide useful information for answering the query. Cuconasu

*Work conducted while FC being a research intern at TII.
†cuconasu@diag.uniroma1.it

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 28022–28036
November 4-9, 2025



(a) (b) (c) (d)

Figure 1: Results of different retrieval pipelines when varying the number k of retrieved passages. We compute the distracting effect on Qwen 2.5 7B.

et al. (2024) divide irrelevant passages as either random, if they are semantically unrelated to the query, or distracting, if they are related to the query but do not contain the answer. They show that while random passages do not affect answer quality, distracting passages do. Jin et al. (2025) show that irrelevant passages returned by strong retrievers are more detrimental than those obtained by weak retrievers. Amiraz et al. (2025) propose a continuous measure of the distracting effect of irrelevant passages and a fine-tuning approach to enhance LLM robustness, similar to strategies in (Lin et al., 2024; Jin et al., 2025; Yoran et al., 2024). To mitigate these challenges, several approaches have emerged to compress or filter retrieved content: Yu et al. (2024) generate sequential reading notes that evaluate document relevance before answer generation, Xu et al. (2024) compress retrieved documents into concise textual summaries using both extractive and abstractive methods with selective augmentation, and Huang et al. (2024) perform sentence-level selection from encoded passages to reduce context length while preserving inference quality.

Positional Bias. Despite advanced positional encoding methods like Alibi (Press et al., 2022) and RoPE (Su et al., 2024), long-context LLMs are typically affected by position bias, i.e., their capability of identifying relevant content depends on its location in the prompt. Liu et al. (2024) discuss the *lost-in-the-middle* effect, where the LLMs tend to ignore information in the middle of the prompt. Hutter et al. (2025) extend this work and demonstrate that different LLMs exhibit distinct positional bias patterns. To mitigate this bias, some solutions propose to fine-tune the LLMs on training data where relevant information is equally distributed across all positions of the prompt (He et al., 2024; An et al., 2024). Other methods modify the attention mechanism of the transformer architecture to remove token-level bias (Leviathan et al., 2025; Ye et al., 2025). Peysakhovich and Lerer (2023) propose a double decoding approach, where in the second decoding step, the passages are re-ordered based on the attention they received in the first step. Jin et al. (2025) re-order the retrieved passages so that top-ranked passages are placed in privileged positions according to the lost-in-the-middle behavior. Zhang et al. (2024) instruct the LLM directly in the prompt to allocate more attention towards a selected segment of the context, aiming to compensate for the shortage of attention. Jiang et al. (2024) mitigates the positional bias by introducing an external module to compress the prompt.

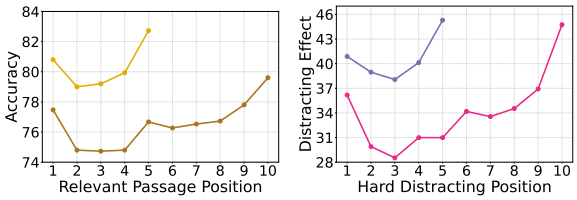
3 Experimental Setup

Benchmarks and Models. We run experiments using the following commonly used public question-answering benchmarks: PopQA (Mallen et al., 2023) and the KILT version (Petroni et al., 2021a) of Natural Questions (NQ) (Kwiatkowski et al., 2019), and TriviaQA (Joshi et al., 2017). From each benchmark, we randomly select two disjoint 500-size samples to run the experiments in Sections 4 and 5, respectively. The results we report in the main paper are averaged across the three datasets[1]. We index the corpus[2] using BM25 (Robertson and Zaragoza, 2009) for sparse retrieval and the *BGE large en v1.5* embedding model (Chen et al., 2024) for dense retrieval. Additionally, we used a re-ranker (RR), namely *BGE reranker v2 m3* (Chen et al., 2024), to rerank the first 25 results from the retriever.

We estimate the performance of the four retrieval pipelines in terms of HITS@k in Fig. 1a, measuring the percentage of times at least a relevant passage is in the top-k retrieved ones, and

[1]Appendix A.2 provides results on each benchmark.

[2]Further details about corpus processing in Appendix A.1.



(a) (b)

Figure 2: Controlled experiments results for Qwen 2.5 7B. **(a)** Average accuracy when rotating a single relevant passage among weak distractors. **(b)** Average distracting effect when rotating a hard distractor among weak distractors. Both exhibit the characteristic U-shaped positional bias pattern.

Precision@k in Fig. 1b, measuring the average percentage of relevant passages in the top-k retrieved ones. Especially when the re-ranker is used, HITS plateaus soon, while Precision keeps decreasing since low-ranked passages are mostly irrelevant. This suggests that using large values of k (e.g., beyond 10) is not worth it, as this would simply add irrelevant passages to the prompt. Therefore, our experiments focus on two reasonable values for k, namely 5 and 10, which provide a good accuracy-latency tradeoff.

As LLMs, we use the instruction-tuned version of Llama 3.2 3B (L3B), Llama 3.1 8B (L8B), Llama 3.3 70B (L70B) (Grattafiori et al., 2024), and Qwen 2.5 7B (Q7B) (Yang et al., 2025), spanning different model sizes and families.

Evaluation Strategy. Following related work (Zheng et al., 2023; Gu et al., 2025; Rahmani et al., 2024), we evaluate passage relevance and answer quality using the LLM-as-a-judge approach. In the former case, we prompt the LLM to assess the relevance of a passage to a question given the ground truth answer, where, following Cuconasu et al. (2024), we consider a passage relevant if and only if it contains the answer to the question, and irrelevant otherwise. In the latter, we prompt the LLM to assess whether the generated response semantically matches the reference answer[3]. We use Claude 3.7 Sonnet via AWS Bedrock as the backbone LLM.[4]

During the experiments, we use the definition of distracting effect introduced by Amiraz et al. (2025). Specifically, their approach consists of prompting an LLM to answer a question q using the information from a passage p or abstain (output "NO-RESPONSE") if the passage does not contain an answer to q. The distracting effect $\mathrm{DE}_q(p)$ of an *irrelevant* passage p for question q is then computed as the probability of the LLM not abstaining:

$$\mathrm{DE}_q(p) = 1 - p^{\mathrm{LLM}}(\text{NO-RESPONSE}|q, p) \quad (1)$$

For each retrieval pipeline, we compute the distracting effect only of irrelevant passages, and assign DE=0 for relevant passages. Fig. 1c reports the DE of the most distracting passage among the top-k positions (MaxDE), while Fig. 1d reports the mean DE considering the top-k positions (MeanDE). Both metrics are averaged across all queries. The MaxDE curves reach very high values, with Table 4 (Appendix) showing that over 60% of queries contain at least one hard distractor (defined as having a DE score greater than 0.8) in the top-10 results from dense retrievers. The MeanDE curves are initially very low, since most of the top retrieved passages are relevant, then increase as more irrelevant passages appear in the prompt, but soon they decrease again. This suggests that highly distracting passages typically appear in top positions, while low-ranked passages have a DE score close to 0. Finally, retrieval pipelines leading to higher HITS and Precision, e.g., when using BGE, also exhibit higher MaxDE and MeanDE curves, revealing a critical aspect: *stronger retrievers increase recall **and** deliver more harmful distractors, making retrieval a double-edged sword.*

Hard Distractor	Relevant Passage Position				
	1	2	3	4	5
None	80.80	79.00	79.20	79.93	82.73
Position 3	75.13	73.80	-	72.40	76.73
Position 5	72.87	71.53	71.60	73.20	-

Table 1: Answer accuracy of Qwen 2.5 7B when rotating a relevant passage in weak distractors only (None), and in weak distractors and a single hard distractor at position 3 or 5.

4 Positional Bias in Controlled Settings

While previous work has established the existence of positional bias in LLMs (Liu et al., 2024; Hsieh et al., 2024; Hutter et al., 2025), these studies typically only analyze the problem from the viewpoint of the relevant passages and completely neglect how the positional bias impacts the effect of distracting passages. In this work, we present the first

[3]Exact prompts are provided in Appendix A.3.

[4]We use Claude 3.7 Sonnet to minimize evaluation errors, though strong open-source models like Llama 3.3 70B Instruct achieve comparable performance (see Appendix A.3.1).

You are given a question and you must respond based on the provided documents. Respond directly without providing any premise or explanation.

Documents:

DE=0.13 **Document[1]** (Title: Bids for the 2024 and 2028 Summer Olympics)(Section: Non-selected bids - 2024 - United States) [...] Following the final presentation, the USOC announced that the United States would bid to host the 2024 Olympic and Paralympic Games, but did not announce which city would bid. On 8 January 2015, the USOC selected Boston to be the candidate city from the United States but on 27 July 2015 Boston's bid was withdrawn and the USOC bid process was reopened. On 1 September 2015 the USOC announced that Los Angeles was chosen for the United States bid for the 2024 Summer Games.

DE=0.00 **Document[2]** (Title: Sports in the United States)(Section: Olympics) [...] The United States hosted both Summer and Winter Games in 1932, and has hosted more Games than any other country – eight times, four times each for the Summer and Winter Games: BULLET::::- the 1904 Summer Olympics in St. Louis, 1932 Summer Olympics and 1984 Summer Olympics in Los Angeles; and the 1996 Summer Olympics in Atlanta; BULLET::::- the 1932 Winter Olympics and 1980 Winter Olympics in Lake Placid, New York; the 1960 Winter Olympics in Squaw Valley, California; and the 2002 Winter Olympics in Salt Lake City, Utah. Los Angeles will host the Summer Olympics for a third time in 2028, marking the ninth time the U.S. hosts the Olympic Games.

DE=0.01 **Document[3]** (Title: 1992 Winter Olympics)(Section: Legacy) The 1992 Olympic Winter Games marked the last time both the Winter and Summer games were held in the same year. The 1992 Olympics also marks the last time France hosted the Olympics. Paris will host the 2024 Summer Olympics.

DE=0.19 **Document[4]** (Title: Sports in Chicago)(Section: Olympic bids) [...] Following Chicago's loss in the race for the 2016 Olympics, the USOC bid for the 2024 Olympics with Los Angeles which result in a deal where Los Angeles secured the right to host the 2028 Summer Olympics. Chicago had previously hosted the 1959 Pan American Games. Chicago was selected to host the 1904 Summer Olympics, but they were transferred to St. Louis to coincide with the Louisiana Purchase Exposition.

DE=0.98 **Document[5]** (Title: Summer Olympic Games)(Section: Hosting) The United States has hosted the Summer Olympic Games four times: the 1904 Games were held in St. Louis, Missouri; the 1932 and 1984 Games were both held in Los Angeles, California; and the 1996 Games were held in Atlanta, Georgia. The 2028 Games in Los Angeles will mark the fifth occasion on which the Summer Games have been hosted by the U.S. [...]

Question: When did the united states host the last olympics?
Answer: The United States hosted the last Summer Olympics in 1996 in Atlanta, Georgia.

Gold Answer: 2002

Figure 3: Example showing how the position of the hard distractor affects Qwen 2.5 7B's response when a relevant passage is fixed in position 2. When a hard distractor (Document 5, DE=0.98) is placed in position 5 (highest distracting effect according to Fig. 2b), the model provides an incorrect answer based on the hard distractor. However, simply moving the hard distractor to position 3 (lowest distracting effect), while maintaining the relevant passage in position 2, results in the model correctly answering "2002".

systematic investigation of the impact of positional bias on distracting passages, analyzing their interactions with relevant content.

For each query, we select the highest-ranked relevant passage obtained by BGE after reranking (BGE+RR). Following Amiraz et al. (2025), we compute the distracting effect for irrelevant passages using Equation 1. We classify passages as "hard distractors" (with DE > 0.8, as previously defined) and "weak distractors" (with DE < 0.2). As an example, Fig. 3 illustrates this distinction: while weak distractors (Documents 1, 3, 4 with DE=0.01-0.19) have minimal impact on the model's reasoning, hard distractors (Document 5 with DE=0.98) can cause the LLM to overlook correct information from relevant passages and generate incorrect answers. Fig. 2 shows results for Qwen 2.5 7B (results for other models and single datasets are given in Appendix B). Fig. 2a displays the characteristic U-shaped accuracy pattern when rotating a single relevant passage among fixed weak distractors[5]. Fig. 2b shows that this positional bias extends to distracting passages, with hard distractors at the beginning or end having significantly higher distracting effect (36-44%) compared to middle slots (28-34%)[6]. This parallel pattern indicates the

[5]We use weak distractors instead of general retrieved irrelevant passages to avoid negative effects from hard distractors.

[6]We calculate the distracting effect using Equation 1 applied to the entire set of passages rather than a single passage.

LLM	Sequential	Inverse	Shuffle	MaxRel	MinDist
Q7B	68.53	71.33	71.00	71.73	70.80
L3B	65.80	68.00	66.73	67.33	66.20
L8B	69.13	69.60	69.87	69.60	69.27
L70B	74.33	74.40	74.60	74.33	75.47

Table 2: Answer accuracy when arranging the top-5 passages retrieved by BGE+RR using different strategies.

Retriever	Sequential	Inverse	Shuffle	MaxRel	MinDist
BGE	68.00	69.00	68.40	68.80	67.47
BGE+RR	68.53	71.33	71.00	71.73	70.80
BM25	51.20	51.27	51.00	51.00	51.00
BM25+RR	59.27	60.20	59.80	59.80	58.80

Table 3: Answer accuracy of Qwen 2.5 7B when arranging with different strategies the top-5 passages retrieved from different retrieval pipelines.

model favors certain positions regardless of passage relevance.

Table 1 further validates this point by showing accuracy when placing a hard distractor at position 3 (lowest DE) versus position 5 (highest DE). We observe an average decrease of about 6 accuracy points compared to using only weak distractors (first row of the table), with a more pronounced drop when the hard distractor occupies position 5. This confirms how positional preference amplifies the negative impact of distracting content.

5 Positional Bias in Real Scenarios

In Section 4, we showed how the answer accuracy can vary up to 5 percentage points in controlled settings, depending on the relevant passage's position. Here, instead, we study the impact of position in real RAG scenarios, i.e., when the LLM prompt contains the top-k ranked passages from the retrieval pipeline. This setting is substantially different from the controlled one shown in Fig. 2a. Indeed, there is no guarantee that a single relevant passage occurs among the top-k ranked passages: there could be none or multiple ones, as well as one or more highly distracting passages. Therefore, we arrange the top-k retrieved passages in the LLM prompt according to the following strategies: (*i*) `Shuffle`: random ordering of passages; (*ii*) `Sequential`: maintaining retrieval ranking order; (*iii*) `Inverse`: inverting the retrieval order, so that according to our LLM prompt template (Fig. 7), the top-1 retrieved passage is the closest to the question; (*iv*) `MaxRelevance`: ranking passages by decreasing positional accuracy estimated during the controlled experiments with the relevant passage[7]. Assuming the retrieval pipeline ranks the passages by decreasing probability of relevance, this strategy maximizes the likelihood of having relevant passages in LLM-favored slots; (*v*) `MinDistraction`: arranging passages by increasing DE order estimated in the controlled setting[8]. Assuming that the retrieval pipeline ranks passages by decreasing DE (as evident in Fig. 1d), this strategy minimizes the likelihood of having highly distracting passages in LLM-favored positions.

Results in Tables 2 and 3 show that the impact of the positional bias in real settings is minor: different passage arrangement strategies lead to very similar results that do not significantly differ from the `Shuffle` baseline[9], regardless of the LLM or the retrieval pipeline. We argue that these results can be explained by the contrastive effect of relevant and highly distracting passages, which, as observed in Fig. 1, tend to both appear in top retrieved passages: for instance, in the `MaxRelevance` strategy, the benefit of placing relevant passages in LLM-favored positions is compensated by the unintended tendency to put in the same slots highly distracting passages.

6 Conclusions

Our work demonstrates that while positional bias exists in current LLMs, its impact is minimal in realistic RAG settings: random ordering of retrieved passages yields statistically equivalent accuracy to more sophisticated reordering strategies. We observed that contemporary retrievers do not merely return some irrelevant passages, they surface passages that degrade answer accuracy in more than 60% of our test questions, turning the retriever itself into a first-order source of error. Thus, attempting to place relevant passages in LLMs' favorable positions may inadvertently prioritize hard distractors over relevant content, counterbalancing the potential benefits of strategic reordering. These findings suggest that future improvements should focus on retrieval quality and LLM distraction robustness rather than passage positioning.

[7]For example, following Fig. 2a for Qwen 2.5 7B with $k = 5$, the estimated order would be *5, 1, 4, 3, 2*.

[8]As an example, following Fig. 2b for Qwen 2.5 7B with $k = 5$ the estimated order would be *3, 2, 4, 1, 5*.

[9]Statistical significance using Wilcoxon test with p=0.05.

Limitations

Our research primarily investigates the factoid question-answering task, though the concept of distracting passages applies to various RAG use cases. Indeed, extending the study to additional tasks, such as multi-hop question answering or fact verification, will provide a more complete picture, but we defer that to future work. Additionally, while we conducted our experiments on English-language benchmarks, the language-agnostic nature of our methodology suggests that the findings would likely generalize to other languages, though formal verification of this hypothesis remains as future work.

Acknowledgments

This work was carried out while Florin Cuconasu was enrolled in the Italian National Doctorate on Artificial Intelligence run by the Sapienza University of Rome. This project has also been supported by PNRR MUR project PE0000013-FAIR and PE00000014-SERICS, funded by European Union – NextGenerationEU, and the NEREO PRIN project funded by the Italian Ministry of Education and Research Grant no. 2022AEFHAZ.

References

Chen Amiraz, Florin Cuconasu, Simone Filice, and Zohar Karnin. 2025. The distracting effect: Understanding irrelevant passages in RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18228–18258, Vienna, Austria. Association for Computational Linguistics.

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2024. Make your LLM fully utilize the context. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

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. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 719–729.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on RAG meeting LLMs: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Junqing He, Kunhao Pan, Xiaoqun Dong, Zhuoyang Song, LiuYiBo LiuYiBo, Qianguosun Qianguosun, Yuxin Liang, Hao Wang, Enming Zhang, and Jiaxing Zhang. 2024. Never lost in the middle: Mastering long-context question answering with position-agnostic decompositional training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13628–13642, Bangkok, Thailand. Association for Computational Linguistics.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. 2024. Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14982–14995, Bangkok, Thailand. Association for Computational Linguistics.

Yufei Huang, Xu Han, and Maosong Sun. 2024. Fast-FiD: Improve inference efficiency of open domain question answering via sentence selection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6262–6276, Bangkok, Thailand. Association for Computational Linguistics.

Jan Hutter, David Rau, Maarten Marx, and Jaap Kamps. 2025. Lost but not only in the middle: Positional bias in retrieval augmented generation. In *Advances in Information Retrieval: 47th European Conference*

on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part I, page 247–261, Berlin, Heidelberg. Springer-Verlag.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2024. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1658–1677, Bangkok, Thailand. Association for Computational Linguistics.

Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan O Arik. 2025. Long-context LLMs meet RAG: Overcoming challenges for long inputs in RAG. In *The Thirteenth International Conference on Learning Representations*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Evgenii Kortukov, Alexander Rubinstein, Elisa Nguyen, and Seong Joon Oh. 2024. Studying large language model behaviors under context-memory conflicts with real documents. In *First Conference on Language Modeling*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1).

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2025. Selective attention improves transformer. In *The Thirteenth International Conference on Learning Representations*.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, Luke Zettlemoyer, and Scott Yih. 2024. RA-DIT: Retrieval-augmented dual instruction tuning. In *The Twelfth International Conference on Learning Representations*.

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 Troy Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021a. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021b. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544.

Alexander Peysakhovich and Adam Lerer. 2023. Attention sorting combats recency bias in long context language models. *Preprint*, arXiv:2310.01427.

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*.

Hossein A. Rahmani, Clemencia Siro, Mohammad Aliannejadi, Nick Craswell, Charles L. A. Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz. 2024. Report on the 1st workshop on large language model for evaluation in information retrieval (llm4eval 2024) at sigir 2024. *Preprint*, arXiv:2408.05388.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomput.*, 568(C).

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *Preprint*, arXiv:2411.04368.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2024. RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation. In *The Twelfth International Conference on Learning Representations*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. 2025. Differential transformer. In *The Thirteenth International Conference on Learning Representations*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations*.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Peixin Cao, Kaixin Ma, Jian Li, Hongwei Wang, and Dong Yu. 2024. Chain-of-note: Enhancing robustness in retrieval-augmented language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14672–14685, Miami, Florida, USA. Association for Computational Linguistics.

Meiru Zhang, Zaiqiao Meng, and Nigel Collier. 2024. Can we instruct LLMs to compensate for position bias? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12545–12556, Miami, Florida, USA. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A Additional Details on the RAG Pipeline

A.1 Corpus and Chunking

We use the KILT knowledge base[10] as corpus for our retrieval. It corresponds to the Wikipedia dump of 01 August 2019. It comprises 5,874,358 Wikipedia articles, which we chunk using SentenceSplitter by LlamaIndex[11] with a chunking size of 256 and no overlap. The splitter tries to segment chunks based on full sentences, avoiding truncations in the middle of a phrase. The chunking phase produced 27,492,989 passages. Then, we index the corpus using Opensearch[12] for sparse retrieval and Pinecone[13] for dense retrieval.

When prompting an LLM with a retrieved passage, we augment it with the title and subsection names from Wikipedia to provide more contextual information to each individual segment (see an example in Fig. 3).

A.2 Additional Retrieval Results

Figures 4 to 6 report the retrieval results of BM25 and BGE with and without re-ranker (RR) on PopQA, NQ, and TriviaQA, respectively.

Moreover, Table 4 shows the percentage of queries that contain at least one hard distractor among the top-k retrieved passages. We define a hard distractor as any irrelevant passage with a distracting effect greater than 0.8.

Retriever	Benchmark	k				
		5	**10**	**15**	**20**	**25**
BGE + RR	NQ	60.60	76.00	81.20	83.00	84.20
	TriviaQA	29.20	44.60	56.20	59.40	61.40
	PopQA	68.40	76.00	79.60	81.20	82.60
	Average	52.73	65.53	72.33	74.53	76.07
BGE	NQ	58.40	73.20	77.20	82.00	84.20
	TriviaQA	28.00	42.60	53.20	59.20	61.40
	PopQA	63.00	72.60	76.00	80.60	82.60
	Average	49.80	62.80	68.80	73.93	76.07
BM25 + RR	NQ	56.60	68.60	71.00	72.20	72.80
	TriviaQA	31.40	42.20	49.80	53.40	54.40
	PopQA	59.80	68.00	71.00	71.80	72.40
	Average	49.27	59.60	63.93	65.80	66.53
BM25	NQ	39.80	55.40	63.20	68.80	72.80
	TriviaQA	25.80	36.60	44.80	50.00	54.40
	PopQA	45.80	59.60	66.60	69.80	72.40
	Average	37.13	50.53	58.20	62.87	66.53

Table 4: Percentage of queries having at least one hard distractor in the top-k retrieved passages.

A.2.1 Multi-Vector Retrieval with ColBERT

We also evaluated ColBERT (Khattab and Zaharia, 2020), a multi-vector retrieval method that performs fine-grained token-level matching between queries and passages. In terms of retrieval quality metrics (HITS@k, Precision@k, MaxDE@k, MeanDE@k), ColBERT exhibits trends similar to BGE (see Fig. 1). For the passage arrangement experiments in Section 5, ColBERT achieved the following accuracies with Qwen 2.5 7B on the top-5 retrieved passages: `Sequential` (67.13), `Inverse`

[10] https://huggingface.co/datasets/facebook/kilt_wikipedia

[11] https://docs.llamaindex.ai/en/stable/api_reference/node_parsers/sentence_splitter/

[12] https://opensearch.org

[13] https://www.pinecone.io/

(67.20), `Shuffle` (67.00), `MaxRelevance` (66.87), and `MinDistraction` (67.07). These results show no statistically significant differences between positional strategies (Wilcoxon test, p=0.05), reinforcing our main conclusion that even sophisticated multi-vector retrievers surface a mixture of relevant and highly distracting passages, thereby mitigating positional bias effects in practical RAG scenarios.

A.3 LLM-as-a-Judge Methodology

A critical aspect of our work is the reliable classification of passages as relevant or irrelevant. We placed particular emphasis on minimizing false negatives, i.e., passages incorrectly labeled as irrelevant despite containing useful information to answer the question. Therefore, we employed a strong LLM, namely Claude 3.7 Sonnet, to judge if a passage is relevant or not. We prompted the LLM to evaluate relevance by considering the question, the passage, the ground truth answers from the dataset, and few-shot examples as demonstrations of relevant and irrelevant passages, with a particular focus on distracting passages. The exact prompt is shown in Fig. 8.

For answer quality evaluation, we prompted the same LLM to assess whether the generated response semantically matches reference answers. This approach prevents penalizing correct answers that use different phrasing than the reference, ensuring our effectiveness metrics genuinely reflect the model's ability to extract and utilize information rather than simply mimic exact answer formats. For example, if the ground truth answer to "What is the population of Yokyo?" is "14 million people", a generated answer like "14 million residents" would be correctly judged as semantically equivalent under our evaluation approach, while it would be considered incorrect under classical exact match metrics. We took inspiration from the OpenAI template used in Wei et al. (2024), with modifications to adapt to our specific task requirements. Fig. 9 provides the exact prompt used for answer quality assessment.

A.3.1 Open-Source Alternative Validation

We also tested whether strong open-source models can serve as reliable alternatives to Claude 3.7 Sonnet for our evaluation tasks. We conducted a comparative analysis using Llama 3.3 70B Instruct on a subset of our data. We sampled 100 queries from each dataset (300 total) and evaluated relevance decisions for the top-5 passages retrieved by BGE using both models, analyzing a total of 1500 passages. For passage relevance assessment, the Cohen's Kappa coefficient between Claude and Llama yielded a score of 0.85, indicating substantial agreement according to standard interpretation guidelines (Landis and Koch, 1977). For answer accuracy evaluation, we had both models assess the correctness of 300 generated answers. The agreement was even stronger, with a Cohen's Kappa score of 0.94, indicating almost perfect agreement. These results suggest that Llama 3.3 70B Instruct can serve as a reliable open-source alternative for both passage relevance assessment and answer quality evaluation in RAG experiments.

B Results for Other LLMs and Single Datasets

In this section, we present detailed results for all LLMs and individual datasets. While the main paper reported results averaged across datasets for space constraints, here we analyze the positional bias effects for each dataset and different LLMs.

B.1 Positional Bias in Controlled Settings

Figures 10 to 13 illustrate the positional bias in controlled settings when rotating either the relevant passage or a hard distractor among weak distractors. The results reveal that each model exhibits its own characteristic positional pattern, confirming findings from Hutter et al. (2025).

Among the LLMs tested, Qwen 2.5 7B demonstrates the most pronounced positional bias (see Fig. 10), while the Llama 3 family appears more resilient to position changes (see Figures 11 to 13). A possible explanation is that these models may have been specifically trained to mitigate the *lost-in-the-middle* effect. Since this problem has become well-documented in the literature (Liu et al., 2024; He et al., 2024; Hsieh et al., 2024), Llama models might incorporate architectural modifications or training techniques designed to maintain robust attention across all positions in the context

LLM	NQ	TriviaQA	PopQA
Q7B	44.20	68.80	20.40
L3B	58.60	68.00	20.60
L8B	67.40	80.80	30.60
L70B	74.60	92.20	49.60

Table 5: Closed-book answer accuracy for different LLMs across the three benchmarks.

LLM	Sequential	Inverse	Shuffle	MaxRel	MinDist
Q7B	70.20	71.00	71.40	71.33	70.33
L3B	64.47	66.47	65.67	65.80	65.73
L8B	68.47	70.80	70.07	68.80	69.00
L70B	75.13	75.00	75.67	76.13	74.33

Table 6: Answer accuracy for different LLMs when arranging the top-10 passages retrieved by BGE+RR using different strategies.

Retriever	Sequential	Inverse	Shuffle	MaxRel	MinDist
BGE	69.33	69.73	68.87	71.00	68.40
BGE+RR	70.20	71.00	71.40	71.33	70.33
BM25	54.73	54.60	55.93	56.07	55.00
BM25+RR	59.93	60.60	60.40	60.73	59.07

Table 7: Answer accuracy of Qwen 2.5 7B when arranging with different strategies the top-10 passages retrieved from different retrieval pipelines.

window, making them less susceptible to passage positioning issues.

In addition, this different behavior from positional bias can be further explained by examining the closed-book effectiveness of these models (Table 5). For the KILT benchmarks, models like Llama 3.3 70B achieve remarkably high closed-book accuracy (74.60 for NQ and 92.20 on TriviaQA), suggesting extensive memorization during pretraining. When LLMs encounter questions they already know the answer to, they tend to rely on their parametric knowledge rather than context, especially when the relevant passage appears in a non-preferential position. This *parametric bias* has been observed by Kortukov et al. (2024), who found that LLMs' factual parametric knowledge can negatively influence their reading abilities and behaviors, leading to a preference for known information over contextual evidence.

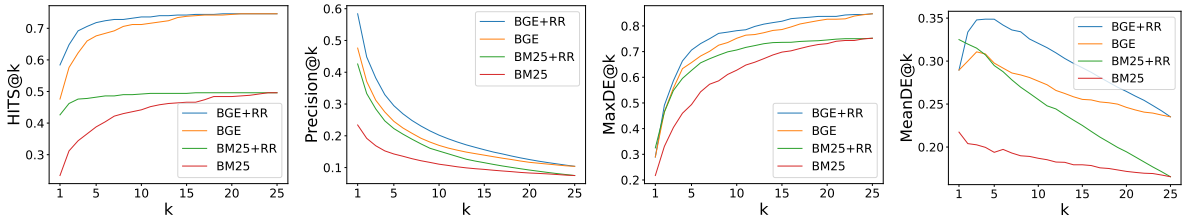
This pattern differs for PopQA, where closed-book accuracy is significantly lower across all models. PopQA contains questions about long-tail entities that are less represented in the models' parametric memory (Mallen et al., 2023), making contextual information more crucial. For smaller models (Llama 3.2 3B, Llama 3.1 8B, and Qwen 2.5 7B), PopQA exhibits stronger positional effects when rotating the relevant passage. The effect is less pronounced in Llama 3.3 70B due to its larger parametric memory that can often recall these long-tail entities.

Regarding distracting effects, when rotating a hard distractor among weak distracting passages, all models generally display the characteristic U-shaped pattern (see Figures 10 to 13), suggesting that distracting effects are more consistent across models and less influenced by parametric knowledge.

B.2 Positional Bias in Real Scenarios

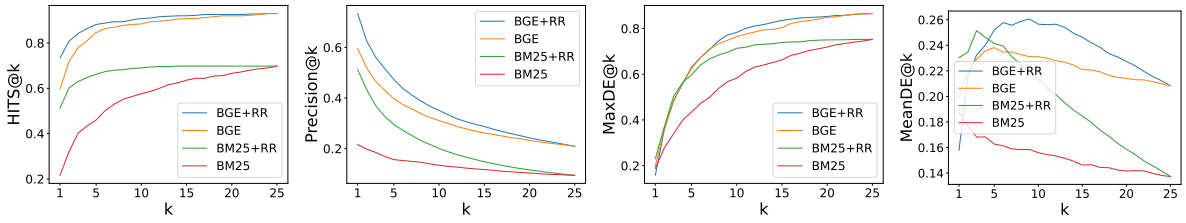
In Section 5, we presented experiments for k=5, showing minimal impact of different passage arrangement strategies on answer accuracy. Here, we expand the analysis to k=10 to investigate whether retrieving more passages might exhibit a more pronounced positional bias effect. Table 6 shows the answer accuracy across different LLMs when arranging the top-10 passages retrieved by BGE+RR using the strategies described in Section 5. Similar to the k=5 case, we observe that the positional bias has a marginal impact on answer accuracy. Across all LLMs, the difference between the best-performing strategy and the `Shuffle` strategy is not statistically significant according to the Wilcoxon test with p-value=0.05. Table 7 presents the results for Qwen 2.5 7B across different retrieval pipelines. We note one exception in the BGE retriever (without re-ranker), where the `MaxRelevance` strategy achieves 71.00 accuracy while `Shuffle` yields 68.87, which is a statistically significant difference. However, this appears to be an isolated case rather than a consistent pattern. This single exception does not contradict the broader statistical trend observed across all other configurations. For weaker retrievers like BM25, the positional ordering has less impact simply because they retrieve fewer relevant passages overall, as shown in Fig. 1a.

In general, these findings with k=10 reinforce our conclusion from the main paper: in realistic RAG settings, the impact of positional bias is minimal compared to its effect in controlled experimental conditions. The interaction between relevant and distracting passages in real retrieval results tends to neutralize potential benefits from strategic passage ordering.



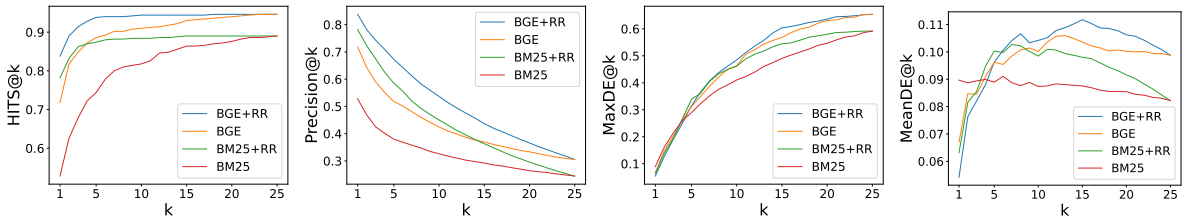
(a) (b) (c) (d)

Figure 4: Results on PopQA of different retrieval pipelines when varying the number k of retrieved passages. We compute the distracting effect on Qwen 2.5 7B.



(a) (b) (c) (d)

Figure 5: Results on NQ of different retrieval pipelines when varying the number k of retrieved passages. We compute the distracting effect on Qwen 2.5 7B.



(a) (b) (c) (d)

Figure 6: Results on TriviaQA of different retrieval pipelines when varying the number k of retrieved passages. We compute the distracting effect on Qwen 2.5 7B.

```
You are given a question and you must respond based on the provided documents. Respond directly
without providing any premise or explanation.

Documents:
<passage>
...
<passage>

Question:
<question>

Answer:
```

Figure 7: Prompt used for response generation.

```
Your job is to look at a question, a list of acceptable answers, and a document, then determine if the document is RELEVANT or IRRELEVANT for
answering the question. Each document may have some metadata information like the title or the section it belongs to. This information may help
you understand the context of the document. We are in a multi-reference setting, which means that there may be multiple correct answers to the
question. The answer list contains all the correct answers.

First, I will give examples of each type, and then you will evaluate a new example.
The following are examples of RELEVANT documents.
```
Question 1: when did korn's follow the leader come out
Acceptable answers list 1: ['August 18 , 1998', 'Summer 1998']
Document 1: (Title: Follow the Leader (Korn album)) Follow the Leader is the third studio album by the American nu metal band Korn . The album was
released on August 18 , 1998 , through Immortal / Epic . This was their first album not produced by Ross Robinson . Instead , it was produced by
Steve Thompson and Toby Wright .

Question 2: who played bobby byrd in get on up
Acceptable answers list 2: ['Nelsan Ellis']
Document 2: (Title: Get on Up (film))(Section: Production - Casting) On August 26, 2013, Universal selected Chadwick Boseman to play the lead role
of James Brown. Boseman did all of his own dancing and some singing. The soundtrack is live recordings of James Brown. On September 17, Universal
announced an open casting call for actors, musicians, and extras for different roles in the biopic, which was held on September 21. On September
30, Taylor cast Viola Davis to play Susie Brown and Octavia Spencer to play Aunt Honey. On October 21, Nelsan Ellis joined the cast of film to
portray Bobby Byrd, Brown's long-time friend.

Question 3: What movie has the song on the road again?
Acceptable answers list 3: ['Honeysuckle Rose']
Document 3: (Title: On the Road Again (Willie Nelson song)) The song , about life on tour , came about when the executive producer of the film
Honeysuckle Rose approached Nelson about writing the song for the film 's soundtrack . '' On the Road Again '' became Nelson 's 9th Country &
Western No. 1 hit overall (6th as a solo recording act) in November 1980 , and became one of Nelson 's most recognizable tunes . In addition ,
the song reached No. 20 on the Billboard Hot 100 , and No. 7 on the Adult Contemporary chart . It was his biggest pop hit to that time and won him
a Grammy Award for Best Country Song a year later .
```
These documents are all RELEVANT because:
    - They contain sufficient information to support at least ONE of the acceptable answers.
    - The information can be found directly or through simple inference.
    - Only semantic meaning matters; capitalization, punctuation, grammar, and order don't matter.

The following are examples of IRRELEVANT documents.
```
Question 1: when did korn's follow the leader come out
Acceptable answers list 1: ['August 18 , 1998', 'Summer 1998']
Document 1: (Title: Korn Discography) Korn's third album marked a significant evolution in their sound and commercial success. The band spent much
of 1998 recording and promoting this album, which would go on to achieve platinum status multiple times. Following their summer tour, they
continued to gain mainstream attention. The album contained several singles that performed well on the charts, including "Got the Life" and "Freak
on a Leash." Reviews were generally positive, with critics noting the band had refined their nu-metal style while maintaining their aggressive
edge.

Question 2: who played bobby byrd in get on up
Acceptable answers list 2: ['Nelsan Ellis']
Document 2: (Title: Get on Up (film))(Section: Critical Reception) Critics particularly praised the casting decisions in "Get on Up," noting the
strong ensemble supporting Chadwick Boseman's portrayal of James Brown. The film's recreation of the dynamic between Brown and his longtime
friend and musical collaborator received significant attention. Several reviewers highlighted the chemistry between the main characters and how
it captured their complex professional and personal relationship spanning decades. The scenes depicting their early musical development were
considered among the film's strongest moments, effectively showing how their partnership shaped the evolution of funk music.

Question 3: What movie has the song on the road again?
Acceptable answers list 3: ['Honeysuckle Rose']
Document 3: (Title: Classic Songs in Films) Many people believe, though it's not actually correct, that Willie Nelson's iconic song 'On The Road
Again' first appeared in the 1980 film 'Smokey and the Bandit II.' Some music historians have suggested that this misconception arose because the
film's themes of truck driving and life on the road seemed to perfectly match the song's message. The song's road-trip vibe made it a natural fit
for many movies, but this particular connection is just a popular misconception.
```
These documents are all IRRELEVANT because:
    - They lack the necessary information to support any of the acceptable answers, even though they may contain some related information.
    - They reference similar themes, keywords, or surrounding context but don't provide the specific answer required.
    - Some contain subtle distractors that seem relevant at first glance but don't actually answer the specific question.

Before making your final evaluation, follow this step-by-step process:
1. Identify the specific information needed to match at least one of the acceptable answers.
2. Carefully search the document for this exact information or information that directly implies it.
3. Check for these common errors:
   - The document contains similar keywords or themes but not the actual answer.
   - The document contains partial information that would need to be combined with external knowledge.
   - The document discusses related topics but doesn't specifically answer the question.
Also note the following things:
- The evaluation should be based ONLY on the specific question and acceptable answers list provided.
- Do not try to generalize or apply your own knowledge beyond the information given in the question, acceptable answers list, and document.
- A document with tangential information about the topic is still IRRELEVANT if it doesn't contain the specific answer.

Here is a new example. Don't apologize or correct yourself if there was a mistake; we are just trying to evaluate the relevance of the document.
```
Question: {question}
Acceptable answers list: {answers}
Document: {document}
```
Evaluate the document for this new question as one of:
A: RELEVANT
B: IRRELEVANT

Return a JSON object with the following format:
{{
    "motivation": "Your concise motivation for the evaluation here. Use maximum 2 sentences.",
    "grade": "A" or "B"
}}
```

Figure 8: Prompt for document relevance assessment using Claude 3.7 Sonnet as judge.

```
Your job is to look at a question, a list of acceptable answers, and a predicted answer, and then assign a grade of either CORRECT or INCORRECT.
We are in a multi-reference setting, which means that there may be multiple correct answers to the question. The answer list contains all the
correct answers.

First, I will give examples of each grade, and then you will grade a new example.

The following are examples of CORRECT predicted answers.
```
Question: What are the names of Barack Obama's children?
Acceptable answers list: ['Malia Obama and Sasha Obama', 'Natasha Marian and Malia Ann']
Predicted answer 1: sasha and malia obama
Preidcted answer 2: Natasha and Malia
Predicted answer 3: most people would say Malia and Sasha, but I'm not sure and would have to double check
Predicted answer 4: Barack Obama has two daughters. Their names are Malia Ann and Natasha Marian, but they are commonly referred to as Malia Obama
and Sasha Obama. Malia was born on July 4, 1998, and Sasha was born on June 10, 2001.
```
These predicted answers are all CORRECT because:
    - They contain all essential information from at least one of the acceptable answers.
    - They do not contain any information that contradicts the acceptable answers.
    - Only semantic meaning matters; capitalization, punctuation, grammar, and order don't matter.
    - Hedging and guessing are permissible, provided that at least one of the acceptable answers is fully included and the response contains no
    incorrect information or contradictions.


The following are examples of INCORRECT predicted answers.
```
Question: What are the names of Barack Obama's children?
Acceptable answers list: ['Malia and Sasha', 'Natasha Marian and Malia Ann']
Predicted answer 1: Malia.
Predicted answer 2: Malia, Sasha, and Susan.
Predicted answer 3: Barack Obama does not have any children.
Predicted answer 4: I think it's either Malia and Sasha. Or it could be Malia and Jackie. Or it could be Joey and Malia.
Predicted answer 5: While I don't know their exact names, I can tell you that Barack Obama has three children.
Predicted answer 6: It's possible you may mean Betsy and Olivia. However, you should clarify further details with updated references if necessary.
Is that the correct answer?
Predicted answer 7: It may be the case that Obama's child is named James. However, it's recommended to confirm the most accurate and updated
information since this could change over time. This model may not always reflect the most current information.
Predicted answer 8: Malia and Sasha are Barack Obama's daughters. Additionally, he adopted a son named Michael in 2018.
```
These predicted answers are all INCORRECT because:
    - A factual statement in the answer contradicts the information in the acceptable answers list. Incorrect statements that have some hedging
    (e.g., "it is possible that", "although i'm not sure, i think") are also considered incorrect.
    In particular:
        - Predicted answer 1: It only mentions one child when both are required.
        - Predicted answers 2, 5, and 8: They include accurate information but also contain incorrect information ("Susan", "three children",
        "adopted a son named Michael").
        - Predicted answer 3: It directly contradicts the acceptable answer.
        - Predicted answer 4: It presents multiple alternatives, some of which are incorrect ("Jackie", "Joey").
        - Predicted answers 6 and 7: They suggest completely incorrect answers, even with hedging ("Betsy and Olivia", "James").


Also note the following things:
- The acceptable answers may contain more information than the question. In such cases, the predicted answer only needs to contain the information
that is in the question.
    - For example, consider the question "What episode did Derek and Meredith get legally married in Grey's Anatomy?" with acceptable answer
    "Season 7, Episode 20: White Wedding". Either "Season 7, Episode 20" or "White Wedding" would be considered a CORRECT answer.
- Do not punish predicted answers if they omit information that would be clearly inferred from the question.
    - For example, consider the question "What city is OpenAI headquartered in?" and the acceptable answer "San Francisco, California". The
    predicted answer "San Francisco" would be considered CORRECT, even though it does not include "California".
    - Consider the question "What award did A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, &
    toxicity win at NAACL '24?", the acceptable answer is "Outstanding Paper Award". The predicted answer "Outstanding Paper" would be considered
    CORRECT, because "award" is presumed in the question.
    - For the question "What is the height of Jason Wei in meters?", the acceptable answer is "1.73 m". The predicted answer "1.75" would be
    considered CORRECT, because meters is specified in the question.
    - For the question "What is the name of Barack Obama's wife?", the acceptable answer is "Michelle Obama". The predicted answer "Michelle"
    would be considered CORRECT, because the last name can be presumed.
- Do not punish for typos in people's name if it's clearly the same name.
    - For example, if the acceptable answer is "Hyung Won Chung", you can consider the following predicted answers as correct: "Hyoong Won
    Choong", "Hyungwon Chung", or "Hyun Won Chung".
- You must attain to the question and acceptable answers list information. Do not rely on your knowledge to grade the answer. If the predicted
answer is correct based on the question and acceptable answer, it should be considered correct, even if you know that the answer is wrong.


Here is a new example. Don't apologize or correct yourself if there was a mistake; we are just trying to grade the answer.
```
Question: {question}
Acceptable answers list: {target}
Predicted answer: {predicted_answer}
```

Grade the predicted answer of this new question as one of:
A: CORRECT
B: INCORRECT

Return a JSON object with the following format:
{{
    "motivation": "Your concise motivation for the grade here. Use maximum 2 sentences.",
    "grade": "A" or "B"
}}
```

Figure 9: Prompt for answer correctness assessment using Claude 3.7 Sonnet as judge.

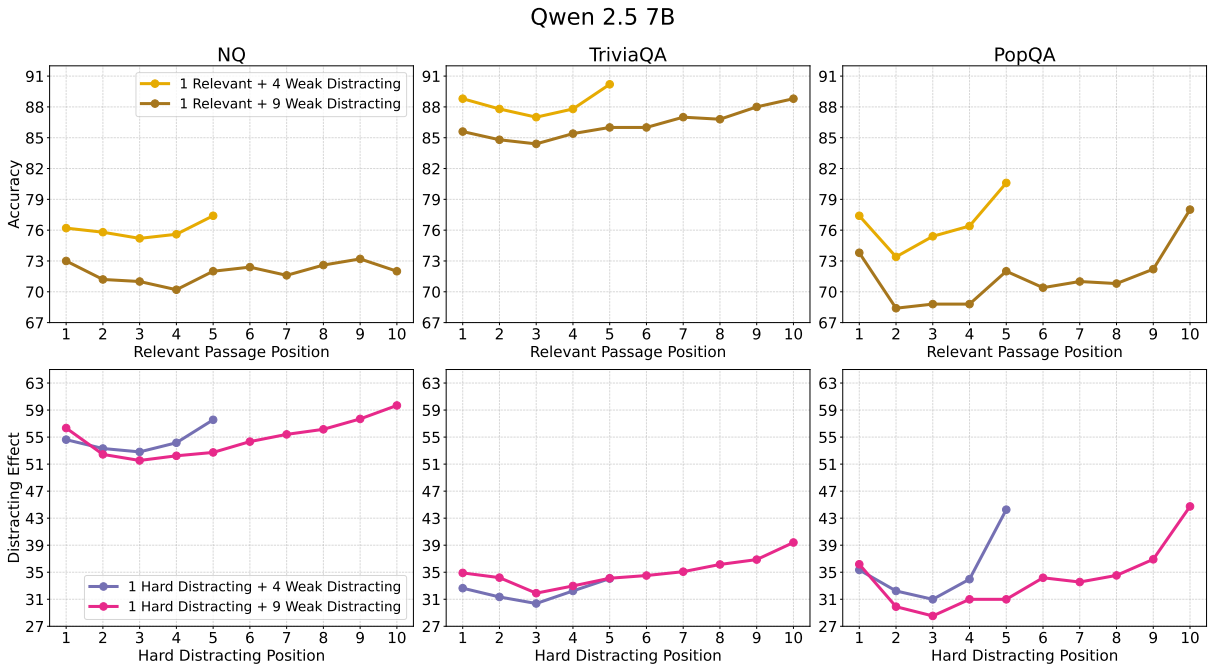


Figure 10: Controlled experiments results for Qwen 2.5 7B across datasets.

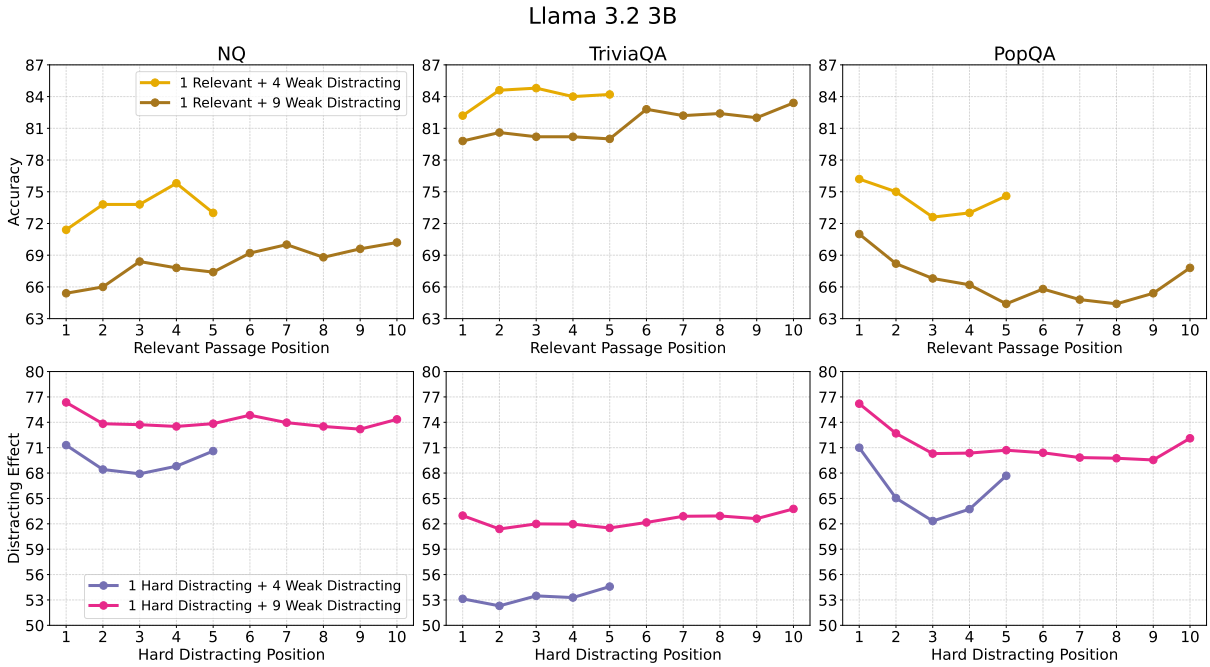


Figure 11: Controlled experiments results for Llama 3.2 3B across datasets.

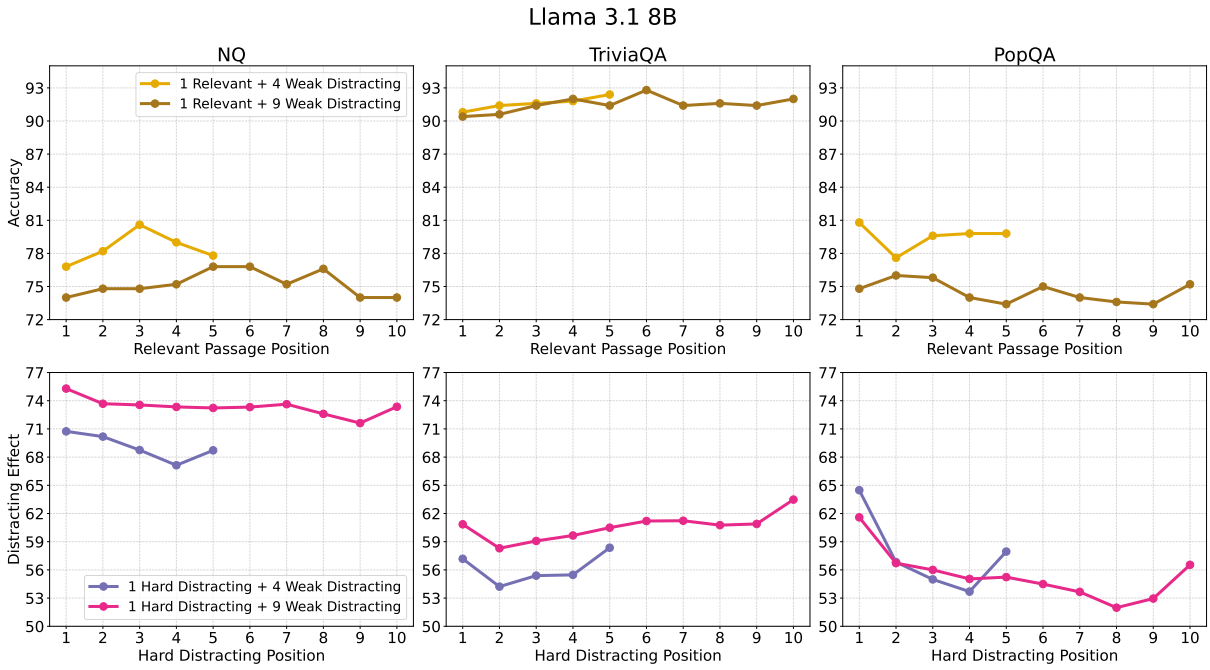


Figure 12: Controlled experiments results for Llama 3.1 8B across datasets.

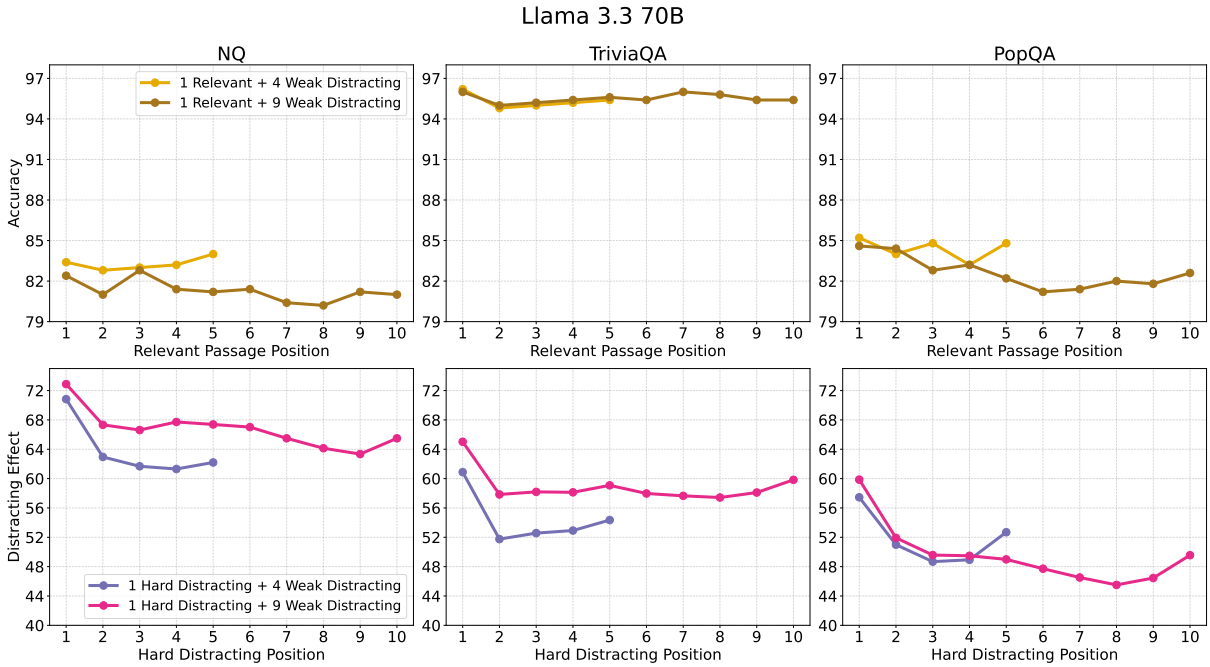


Figure 13: Controlled experiments results for Llama 3.3 70B across datasets.