Can We Instruct LLMs to Compensate for Position Bias?

Anonymous ACL submission

Abstract

Position bias in LLMs leads to difficulty in accessing information retrieved from the retriever, thus downgrading the effectiveness of RAG approaches in open-question answering. Recent studies reveal that this bias is related to disproportional attention across the context. In this work, we examine directing the LLMs to allocate more attention towards a selected segment of the context through prompting, such that the shortage of attention can be compensated. We find that language models do not have relative position awareness of the context but can be directed by promoting instruction with an exact document index. Our analysis contributes to a deeper understanding of position bias in LLMs and provides a pathway to mitigate this bias by instruction, thus benefiting LLMs in locating and utilizing relevant information from retrieved documents in RAG applications.

1 Introduction

Retrieval-Augmented Generation (RAG) is an established method for enabling continuous knowledge updates (Wu et al., 2024; Gao et al., 2023; Chu et al., 2024; Lewis et al., 2020) and reducing hallucination (Ji et al., 2023; Zhang et al., 2023) through retrieving and adding relevant documents to the prompt of LLMs (Glass et al., 2022; Xu et al., 2024). However, recent research has discovered that increasing the number of documents in the context may distract the model and degrade performance (Weller et al., 2024; Oh and Thorne, 2023), even when they contain accurate and relevant information (Sauchuk et al., 2022).

Indeed, increasing evidence indicates that LLMs struggle to use context effectively due to position bias (Xiao et al., 2024; Liu et al., 2023; Zheng et al., 2023; Qin et al., 2023) that the models favor the beginning or end text within the context (Liu et al., 2024a), leading to the "lost-in-the-middle" problem. For example, Figure 1 illustrates this problem in the RAG pipeline for the open question answering task, where multiple retrieved documents are added to the prompt. By grouping and averaging the attention scores of tokens across the 3 retrieved documents, we observe that the second document consistently receives less attention scores, irrespective of the gold document's position, which aligns with previous works (Chen et al., 2023; Zhang et al., 2024; He et al., 2024). This bias can lead to incorrect answers when the gold document is in the middle.

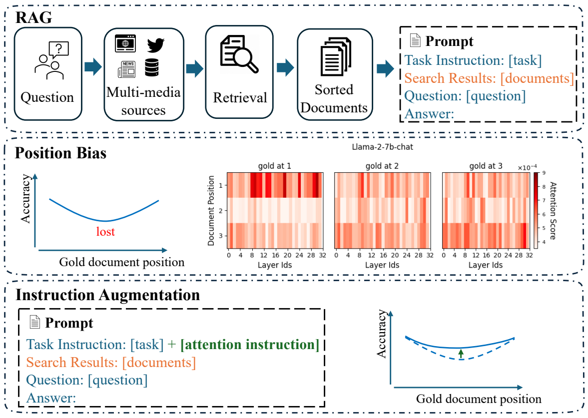

Figure 1: **Top:** An example of RAG for open question answering, where the prompt contains the sorted documents. **Middle:** The position bias (i.e. lost in the middle) can be visualized by attention score, which shows a significant drop in the middle wherever the gold answer is placed. **Bottom:** We solve this by augmenting the prompt with an attention instruction.

To address position bias, many researchers have explored either finetuning (He et al., 2023; An et al., 2024; Fu et al., 2024; Wang et al., 2023) or modifying position embeddings (Chen et al., 2023; He et al., 2024; Zhang et al., 2024)[1]. However, finetuning-based approaches lack adaptability and require additional computation, whereas embedding-based approaches require multiple rounds of inferencing or hyperparameter search,

[1] Details of the related work can be found in Appendix A.1.

which is inefficient.

In this study, we focus on instructing LLMs to attend to specific positions within the context, thereby compensating for position bias. In particular, we design two types of `attention instructions` that instruct LLMs to adjust their attention using either relative position words or absolute document indexes. We conduct comprehensive experiments with these two types of attention instructions on five open-sourced LLMs based on the multi-document question-answering task. Our investigation focuses on the feasibility of mitigating position bias in LLMs through attention instructions.

In summary, our findings are as follows:

- Our experimental results indicate that language models lack an understanding of positional concepts and therefore fail to follow the relative attention instruction.
- Our investigation on absolute attention instruction shows evidence that the attention of LLMs to a segment within the context can be enhanced semantically.
- We illustrate that relative regional attention control can be achieved by attaching the same index to multiple documents.

2 Experimental Setup

We design `attention instructions`, a two-sentence prompt that guides LLMs to focus on a selected segment, thereby preventing the overlooking of crucial information. To test the effectiveness of the attention instructions, we conduct a series of experiments on the multi-document question answering (MDQA) task (Singh et al., 2021) under the setting that only one document contains the gold answer, namely the gold document. The position of the gold document is referred to as the *gold document position*. By controlling the gold document position and attention segment specified in the instructions, we aim to evaluate the LLMs' ability to follow attention instructions accurately. An overview of the input prompt and some toy examples can be seen in Figure 2.

2.1 Attention instruction

The attention instruction is a two-sentence instruction that aims to guide the model to focus on a positional segment of the search results. Hereafter we refer to the phrase representing the position of segment in instructions as *attention segment phrase*. The first sentence explicitly informs the model where the answer is located, while the second sentence directs the model to use that segment as the main reference when answering the question. To investigate the effectiveness of attention instructions in mitigating position bias, we explore relative attention instruction and absolute attention instruction. The details are as follows:

- **Relative Attention Instruction**: We use the phrase "`{position} part`" to guide the model's focus on a positional segment of the search results. The position words *beginning*, *midsection*, and *tail* are used to virtually split the search results into three parts.
- **Absolute Attention Instruction**: We use the document indexes as the segment phrase in attention instruction. There are two types of indexes, ID-Index (e.g. 1, 2, 3) and Position-Index (e.g. relative position represented by the position words listed above). For ID-index, we use "`document [{ID}]`". For the position-index, we directly use the position words as the attention segment phrase.

Figure 2 shows the prompt structure after adding the attention instructions, as well as the illustrations of No-Index, ID-Index and Position-Index.

2.2 Datasets and models

We use the dataset created by Liu et al. (2024a), which contains 2,655 data samples and each example in the dataset consists of a tuple with question, answer, gold document, distractor documents, where the distractor documents are relevant to the questions but do not contain the corresponding answer[2]. We use accuracy as our evaluation metric, considering an answer correct if the gold answer exists in the generated output.

We experiment with five state-of-the-art open-sourced models that are instruction-tuned including Llama-2-7b-chat (Touvron et al., 2023), Meta-Llama-3-8B (Meta AI Research, 2023), Tulu-2-7b (Ivison et al., 2023), Mistral-7B-Instruct-v0.1 and Mistral-7B-Instruct-v0.2 (Jiang et al., 2023).

3 Result and Analysis

Probe the relative position awareness of LLMs with relative attention instructions As described in §2.1, we virtually split the search results

[2] Details of the construction of dataset can be found in Appendix A.3.

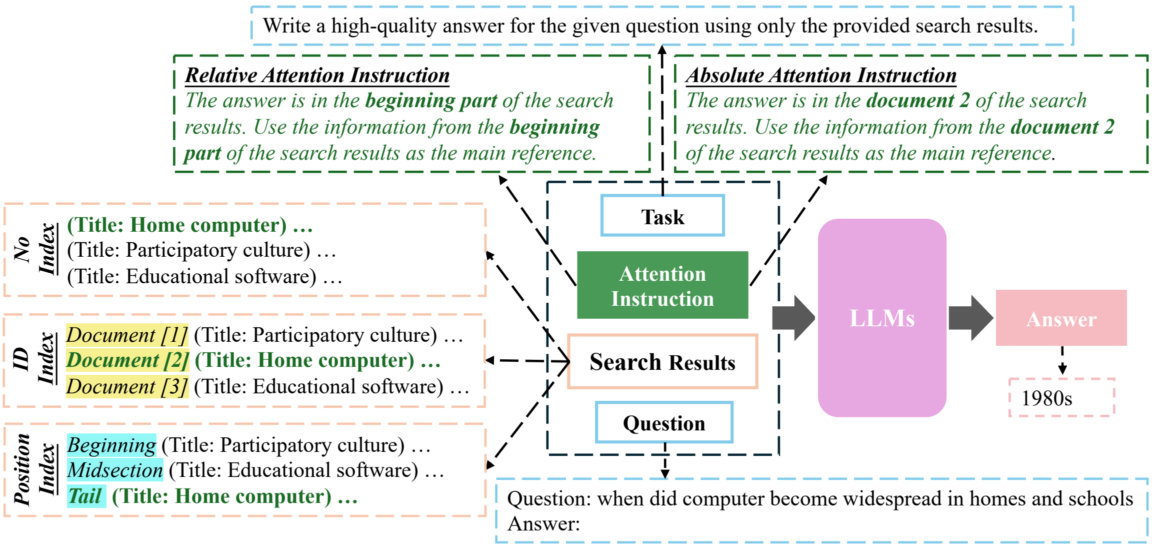

Figure 2: Prompt structure. The top two □ boxes show the two types of attention instructions, where the attention segment phrase is marked in ***bold***. Three index types for documents (highlighted in yellow for ID-index and cyan for position-index) are shown in the left □ boxes, with the **gold document** shown in different positions.

into three parts and represent these parts with relative positions words *beginning*, *midsection*, and *tail*. By placing the gold document at different positions among all the documents and refer to each position in the relative attention instruction, we create a 3x3 accuracy heatmap for each model. The heatmaps' y-axis represents the gold document position, while the x-axis represents the selected attention segment. It is worth noting that diagonal cells in the heatmap reflect instances where the attended segments align with the positions of the gold documents.

We present the accuracy heatmaps of Meta-Llama-3-8B, and Mistral-7B-Instruct-v0.2 with 3 documents in Figure 3. In general, the top row in each heatmap outperforms the other rows, which is consistent with the findings of Liu et al. (2024a), indicating that the model is biased to the beginning (3-5% higher than the midsection and tail). There is no significant improvement observed in diagonal cells across different positions on both models, which suggests that LLMs do not follow the relative attention instruction effectively and reveals a lack of relative position awareness among LLMs[3].

Instruct LLMs with document ID-Index and absolute attention instruction As illustrated in Figure 4, when the document ID is used as a index to each document and a reference in the absolute attention instruction, the models' performance on the diagonals across all models are boosted, especially Llama-2-7b-chat (4% to 10% ↑). Conversely, when

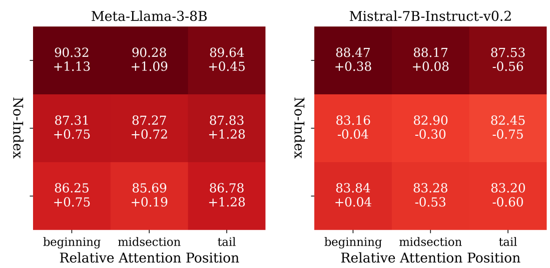

Figure 3: Accuracy heatmaps of Meta-Llama-3-8B and Mistral-7B-Instruct-v0.2 when using relative attention instruction in No-Index setting. In each cell of the heatmaps, the accuracy value is shown in % and the $+-$ indicates the performance difference compared to without using attention instruction. The darker the color of the cell, the higher the accuracy.

LLMs are instructed to focus on distractor documents, the performance drops significantly (e.g., 25% ↓ for Llama-2-7b-chat when the gold document is at the beginning). This suggests that absolute attention instructions enable LLMs to focus on specific documents, mitigating the position bias[4].

When comparing cross models, the Llama-2-7b-chat model is more sensitive to attention instructions. Meta-Llama-3-8B exhibits better instruction-following ability than Mistral-7B-Instruct-v0.2, despite having similar absolute accuracy. Tulu-2-7b, a finetuned Llama-2 model, is less sensitive to absolute attention instruction and maintains robustness when guided to attend to distractor documents compared to Llama-2-7b-chat, possibly due to its ex-

[3] The results of all models in both 3-document and 9-document setting in Appendix A.5.1 support our conclusion.

[4] Full results can be found in Appendix A.5.2. The results present the effectiveness of absolute attention instruction.

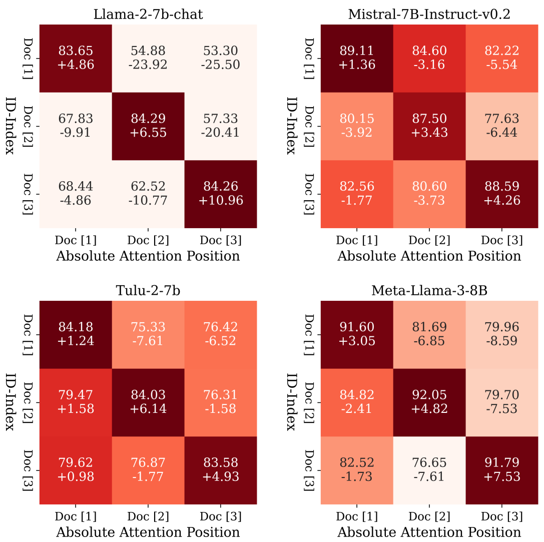

Figure 4: Results of Llama-2-7b-chat, Mistral-7B-Instruct-v0.2, Tulu-2-7b, Meta-Llama-3-8B using absolute attention instruction with relative numerical IDs as document indexes.

tended context window (from 4096 to 8192 tokens) and new data mixture used during finetuning.

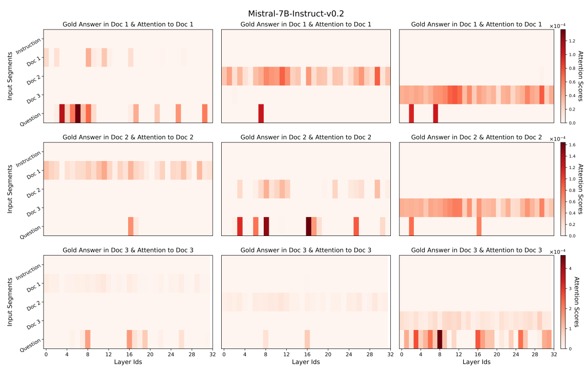

Figure 5: The attention score heatmaps of Mistral-7B-Instruct-v0.2 using absolute attention instruction with document ID index.

Figure 5 visualizes the attention scores for each segment of the prompt, arranged in the same way as the accuracy heatmaps, to investigate the impact of attention instructions on attention score distribution. Each subplot represents a pair of gold document positions and attention segments. The color bar starts at 0, and white areas may have reduced or unchanged attention scores. When the model is instructed to focus on a specific document based on its ID, the average attention score of the tokens in that document increases, regardless of the gold document position. When the attention segment matches the gold document position, the attention to the question also improves, suggesting that attention instructions encourage the model to consider the question more when seeking the answer. Comparing across layers, we observe that the front layers are more sensitive to absolute attention instructions.

Instruct LLMs to attend to relative positions with absolute attention instruction To investigate the feasibility of achieving regional attention control through absolute attention instructions, we conduct experiments with a 9-document setting[5], where three documents are grouped together and assigned the same position index. We refer to relative positions in the attention instruction and the results presented in Figure 6 reveal a subtle but distinct diagonal pattern, indicating improved performance when models are instructed to attend to the region containing the gold document, and deteriorated performance in mismatched cases. The results demonstrate that absolute attention instructions can effectively guide LLMs to focus on specific regions of the search results by assigning the same index to multiple documents, thus enabling regional attention control.

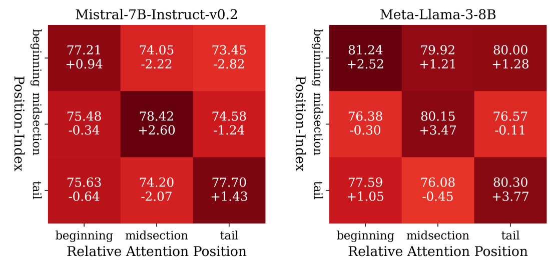

Figure 6: 9-document results of Mistral-7B-Instruct-v0.2 and Meta-Llama-3-8B using absolute attention instruction with position-index.

4 Conclusion and Future Work

We empirically study how sensitive LLMs are to attention instructions via a series of systematic experiments. We find that LLMs can be prompted to pay more attention to a document or region through direct indexing. However, we also find that models are not capable of locating a document or a region in the context based on its relative positions. Our results and analyses provide new insights into solving the position bias through semantic instructions and a potential pathway to achieve more effective RAG by distributing attention based on relevance scores or source information confidence.

[5] We present the result in a 9-document setting since its position bias is more severe than 3-document. Appendix A.2 shows its prompt. The results of the 3-document setting are shown in Appendix A.5.3, which leads to the same conclusion.

5 Limitations

Our study has several limitations that should be acknowledged. First, we limited the search results to include only one document containing the gold answer, while real-world scenarios may involve multiple documents with correct or partially correct answers and conflicting information. Moreover, the gold document position is unknown in real-world scenarios, requiring a pre-identification of the attention position when implementing attention instructions in RAG applications. Future research could explore the effectiveness of attention instructions in these more complex settings. Second, due to computational resource limitations, we experimented with a maximum of 9 documents and tested models with sizes ranging from 7B to 8B, leaving the exploration of larger contexts and models for future work. Finally, we focused on the correlation between semantic prompts and attention values, and did not investigate closed-source language models. Future research could expand the scope by examining the attention instruction following capabilities of these models. Addressing these limitations and exploring attention instructions in more diverse settings will further enhance our understanding of their potential and guide the development of more effective RAG models.

6 Ethics Statement

In preparing and submitting this research paper, we affirm that our work adheres to the highest ethical standards and is devoid of any ethical issues. The study did not involve any human subjects or sensitive data, and all models and datasets used are publicly available. We acknowledge the potential risks associated with large language models and have focused our research on understanding their attention mechanisms to contribute to the development of more transparent and controllable models.

References

Kosuke Akimoto, Kunihiro Takeoka, and Masafumi Oyamada. 2023. Context quality matters in training fusion-in-decoder for extractive open-domain question answering. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, and Jian-Guang Lou. 2024. Make your llm fully utilize the context. *ArXiv preprint*, abs/2404.16811.

Yuhan Chen, Ang Lv, Ting-En Lin, Changyu Chen, Yuchuan Wu, Fei Huang, Yongbin Li, and Rui Yan. 2023. Fortify the shortest stave in attention: Enhancing context awareness of large language models for effective tool use. *ArXiv preprint*, abs/2312.04455.

Zhendong Chu, Zichao Wang, Ruiyi Zhang, Yangfeng Ji, Hongning Wang, and Tong Sun. 2024. Improve temporal awareness of llms for sequential recommendation. *ArXiv preprint*, abs/2405.02778.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. 2024. Data engineering for scaling language models to 128k context. *ArXiv preprint*, abs/2402.10171.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *ArXiv preprint*, abs/2312.10997.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2g: Retrieve, rerank, generate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2701–2715.

Junqing He, Kunhao Pan, Xiaoqun Dong, Zhuoyang Song, Yibo Liu, Yuxin Liang, Hao Wang, Qianguo Sun, Songxin Zhang, Zejian Xie, et al. 2023. Never lost in the middle: Improving large language models via attention strengthening question answering. *ArXiv preprint*, abs/2311.09198.

Zhiyuan He, Huiqiang Jiang, Zilong Wang, Yuqing Yang, Luna Qiu, and Lili Qiu. 2024. Position engineering: Boosting large language models through positional information manipulation. *ArXiv preprint*, abs/2404.11216.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A Smith, Iz Beltagy, et al. 2023. Camels in a changing climate: Enhancing lm adaptation with tulu 2. *ArXiv preprint*, abs/2311.10702.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *ArXiv preprint*, abs/2310.06825.

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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024a. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Yijin Liu, Xianfeng Zeng, Fandong Meng, and Jie Zhou. 2023. Instruction position matters in sequence generation with large language models. *ArXiv preprint*, abs/2308.12097.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulic, Anna Korhonen, and Nigel Collier. 2024b. Aligning with human judgement: The role of pairwise preference in large language model evaluators. *arXiv preprint arXiv:2403.16950*.

Zhongkun Liu, Zheng Chen, Mengqi Zhang, Zhaochun Ren, Zhumin Chen, and Pengjie Ren. 2024c. Zero-shot position debiasing for large language models. *ArXiv preprint*, abs/2401.01218.

Meta AI Research. 2023. Meta LLaMA 3: Improving Instruction Following and Few-Shot Learning. Accessed: 30 April 2024.

Philhoon Oh and James Thorne. 2023. Detrimental contexts in open-domain question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11589–11605.

Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. In *Automated Knowledge Base Construction*.

Alexander Peysakhovich and Adam Lerer. 2023. Attention sorting combats recency bias in long context language models. *ArXiv preprint*, abs/2310.01427.

Guanghui Qin, Yukun Feng, and Benjamin Van Durme. 2023. The nlp task effectiveness of long-range transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3774–3790.

Artsiom Sauchuk, James Thorne, Alon Halevy, Nicola Tonellotto, and Fabrizio Silvestri. 2022. On the role of relevance in natural language processing tasks. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1785–1789.

Devendra Singh, Siva Reddy, Will Hamilton, Chris Dyer, and Dani Yogatama. 2021. End-to-end training of multi-document reader and retriever for open-domain question answering. volume 34, pages 25968–25981.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint*, abs/2307.09288.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2023. Augmenting language models with long-term memory. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Orion Weller, Aleem Khan, Nathaniel Weir, Dawn Lawrie, and Benjamin Van Durme. 2024. Defending against poisoning attacks in open-domain question answering. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 402–417.

Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Cheng Niu, Randy Zhong, Juntong Song, and Tong Zhang. 2024. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. *ArXiv preprint*, abs/2401.00396.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*.

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

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *ArXiv preprint*, abs/2309.01219.

Zhenyu Zhang, Runjin Chen, Shiwei Liu, Zhewei Yao, Olatunji Ruwase, Beidi Chen, Xiaoxia Wu, and Zhangyang Wang. 2024. Found in the middle: How language models use long contexts better via plug-and-play positional encoding. *ArXiv preprint*, abs/2403.04797.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. On large language models' selection bias in multi-choice questions. *ArXiv preprint*, abs/2309.03882.

Han Zhou, Xingchen Wan, Lev Proleev, Diana Mincu, Jilin Chen, Katherine A Heller, and Subhrajit Roy. 2024. Batch calibration: Rethinking calibration for in-context learning and prompt engineering. In *The Twelfth International Conference on Learning Representations*.

A Appendix

A.1 Related work

Retrieval Augmented Generation Petroni et al. (2020) were the first to apply RAG with pretrained language models on unsupervised question answering. Lewis et al. (2020) originated the extractive open-domain question answering with retrieval augmentation. While the external knowledge and information provide solutions to open-domain question answering (Izacard and Grave, 2021), LLMs still have difficulty in leveraging the retrieved passages effectively (Sauchuk et al., 2022; Oh and Thorne, 2023). Despite the conflicting misinformation and detrimental passages (Weller et al., 2024; Oh and Thorne, 2023), disproportional attention distribution towards passages also introduces challenge (Akimoto et al., 2023). This work considers the RAG setting, assuming the search results are given.

Position bias in LLMs Recent studies have demonstrated that the position of instruction (Liu et al., 2023) and the order of answer choices (Zheng et al., 2023) within the context can affect the performance and generation of LLMs. LLMs also have primary bias and recency bias in which the attention scores are biased towards initial tokens and the context in the end, regardless of their semantic relevance to the task (Xiao et al., 2024; Qin et al., 2023). Liu et al. (2024a) investigated the long-context reasoning of LLMs and noted the challenge that the information in the middle is likely to be overlooked.

Addressing position bias through context reordering and finetuning Some researchers propose mitigating position bias by reordering the context based on relevance (Wang et al., 2023; Peysakhovich and Lerer, 2023; Liu et al., 2024b). However, these explicitly designed orders may not always work as expected (Liu et al., 2024a). Others suggest addressing position bias through continual finetuning of LLMs (He et al., 2023; An et al., 2024; Fu et al., 2024; Liu et al., 2024c). These methods aim to strengthen attention over all parts of the context or scale up LLMs' context window length without losing information accessing capability, but they require processing training data and additional finetuning, which can be computationally expensive.

Addressing position bias through position embedding modification and logits calibration Chen et al. (2023); He et al. (2024); Zhang et al. (2024) suggest that RoPE (Rotary Position Embedding) introduces long-term attention decay and propose modifying the position embeddings to address position bias. Chen et al. (2023) merges the attention of multiple parallel runs with different RoPE bases, while Zhang et al. (2024) re-scales the position indices to smaller values. He et al. (2024) adjusts the attention scores by adding placeholder tokens between different segments to mitigate the effect of instruction on the adjacent document. However, these approaches either require parallel runs or hyperparameter tuning, introducing additional computational overhead. Alternatively, Zhou et al. (2024) introduced Batch Calibration (BC), a zero-shot and inference-only calibration method that estimates the contextual bias by marginalizing the LLM scores in the batched input, addressing biases in LLMs without modifying position embeddings. In contrast to these approaches, we focus on leveraging the instruction-following capability of LLMs to achieve fine-grained usage of different documents and investigate the implicit correlation between semantic attention and the attention scores of LLMs.

A.2 Prompt template for 9-document Position-Index

Prompt template

Write a high-quality answer for the given question using only the provided search results. The answer is in the ***beginning*** of the search results. Use the information from the ***beginning*** of the search results as the main reference.

beginning (Title: Participatory culture)...
***beginning* (Title: Home computer)...**
beginning (Title: Educational software)...
midsection (Title: Ronald Anderson)...
midsection (Title: Computers in the classroom)...
midsection (Title: Warez)...
tail (Title: History of computer hardware in Yugoslavia)...
tail (Title: Altair 8800)...
tail (Title: Steven Paul Rudolph)...

Question: when did computer become widespread in homes and schools
Answer:

Figure 7: Prompt template for combining absolute attention instruction with position indexes.

A.3 Dataset Details

The question, answer and gold document are from NaturalQuestion-Open dataset (Kwiatkowski et al.,

2019) and $n-1$ distractor documents that are relevant but do not contain the answer are retrieved using a retrieval system (Contriever, finetuned on MS-MARCO; (Izacard et al., 2022)). To ensure consistency and control input length, all documents are chunked to a maximum of 100 tokens.

A.4 Attention Scores Case Study

Figure 8 presents an example where the model initially struggles to answer correctly without additional guidance but provides the correct answer after using an absolute attention instruction.

In this example, the gold document is placed in the middle, and we use absolute attention instruction to guide the model to pay more attention to document 2. By plotting the attention score difference after applying the attention instruction, we observe a clear increase in the attention scores of document 2. The increased attention scores on document 2 suggest that self-attention affects answer prediction and that guiding the language model through absolute attention instructions can help address challenging questions where the crucial information required for answering the question is harder to find.

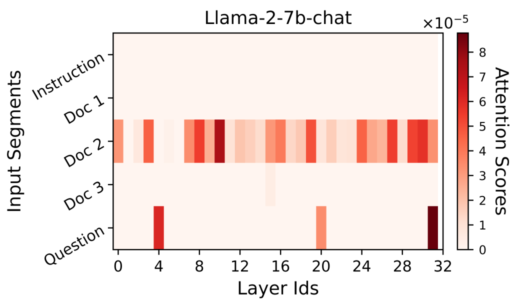

Figure 8: Case study: the attention score of an example that answers correctly after using attention instruction.

A.5 Additional Results and Analysis

This section presents additional results and analysis of the models in different instruction and index settings to further support our findings and conclusions in §4. Due to space constraints, the main content primarily includes results of Llama-2-7b-chat, Meta-Llama-3-8B, and Mistral-7B-Instruct-v0.2 under specific settings. Here, we provide a more comprehensive set of results for all five models in both 3-document and 9-document settings.

A.5.1 Relative Attention Instructions with No-Index. We show the accuracy heatmaps of all five models using relative attention instructions and no index added to the documents in both 3-document (Figure 9) and 9-document (Figure 10) settings. The results confirm that the lack of significant differences after using relative attention instructions is consistent across all models, reinforcing the finding that LLMs do not have relative position awareness and cannot effectively follow relative attention instructions.

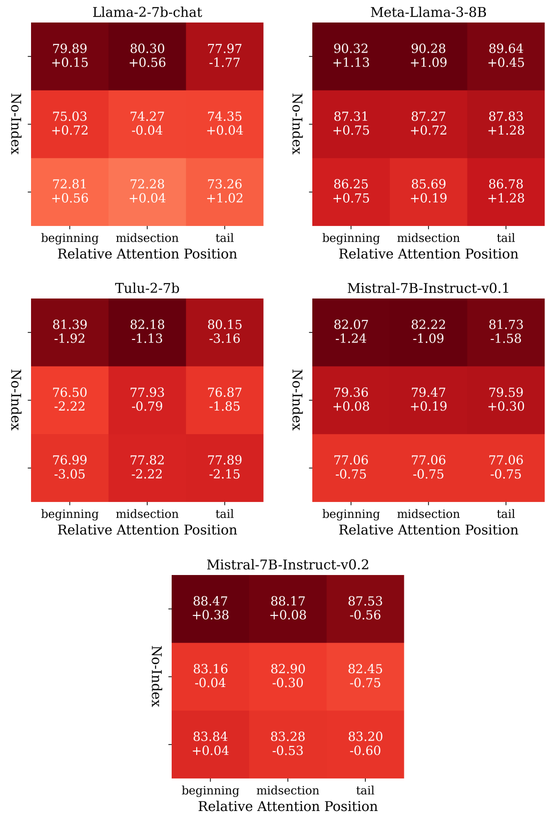

Figure 9: 3-document: relative attention instruction under no-index setting.

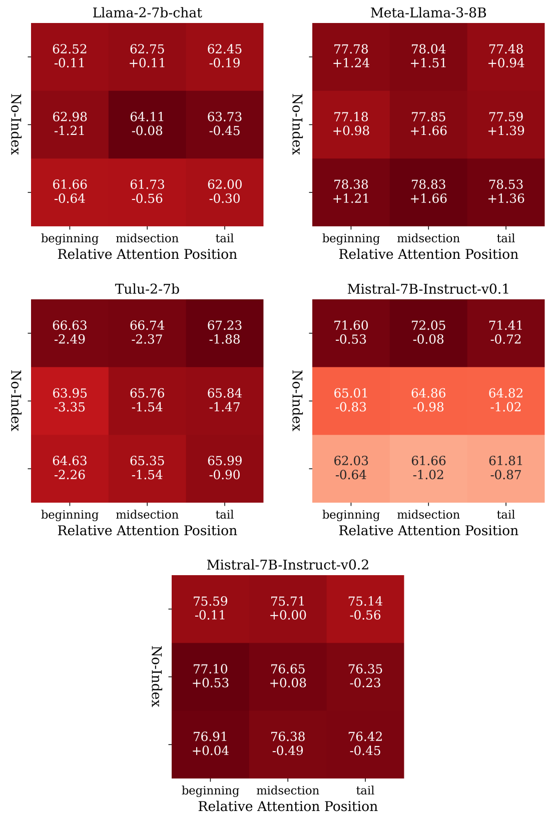

Figure 10: 9-document: relative attention instruction under no-index setting.

A.5.2 Absolute Attention Instructions with ID-Index. We investigate the effectiveness of absolute attention instructions in both 3-document and 9-document settings with ascending document ID indexes for all five models (Figure 11 and Figure 12). The results validate the generalized applicability of absolute attention instructions, demonstrating that despite the increasing number of distractor documents, referencing the exact document ID of the gold document boosts model performance. Comparing the 3-document and 9-document results of Llama-2-7b-chat and Mistral-7B-Instruct-v0.2 reveals that the significance of attention instructions is also influenced by the document's relative position (e.g., beginning or tail). In contrast, the influence of attention instructions on Tulu-2-7b and Meta-Llama-3-8B is less sensitive to document position.

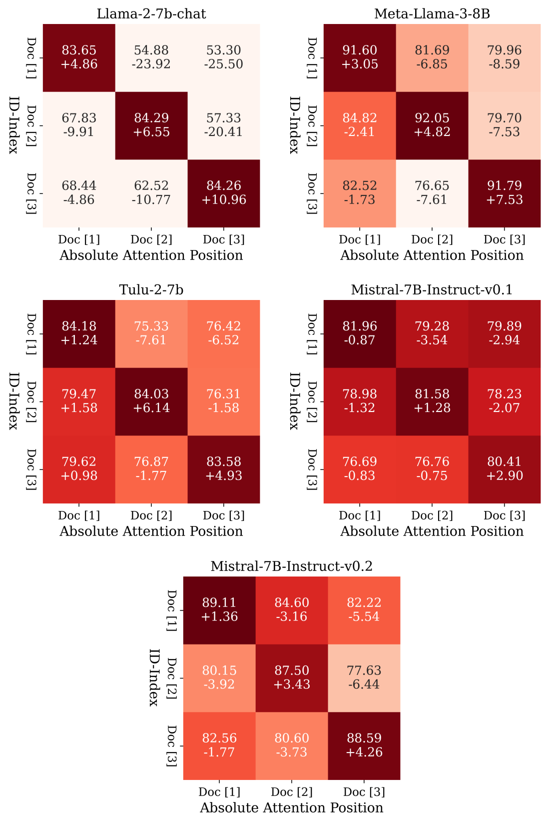

Figure 11: 3-document: absolute attention instruction under ID-index setting.

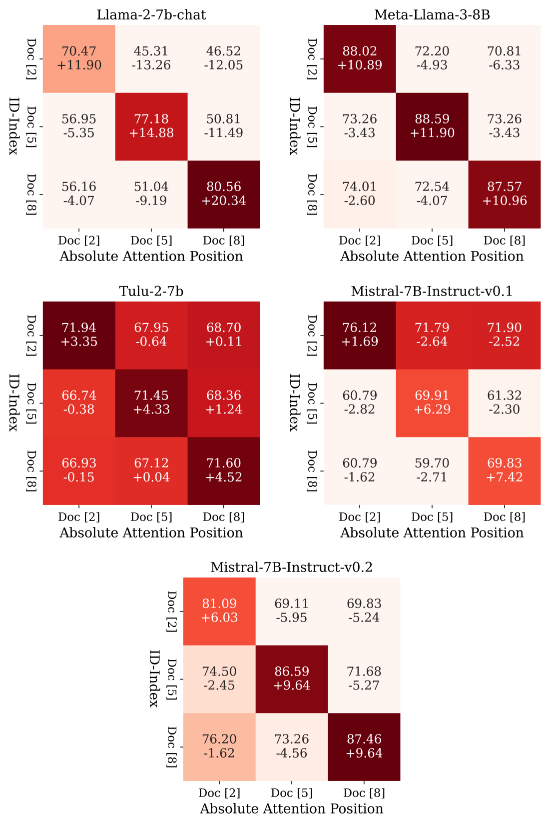

Figure 12: 9-document: absolute attention instruction under ID-index setting.

A.5.3 Positional Control Using Absolute Attention Instructions with Position-Index. To complement the results for RQ3, we present the results of using absolute attention instructions with position-index for all five models in both 3-document (Figure 13) and 9-document (Figure 14) settings. The clear diagonal pattern in the accuracy heatmaps for both settings supports our finding that position words can serve as effective indexes for documents in each part of the search results, enabling regional control through attention instructions. The 3-document setting results (Figure 13) show that using position-index leads to improved performance when the attention instruction matches the gold document's position, consistent with the findings in the main content. The 9-document setting results (Figure 14) further demonstrate the effectiveness of using position-index for regional control, as the models exhibit improved performance when instructed to attend to the region containing the gold document. These additional results and analysis emphasize the consistency of our findings across different models, instruction types, and index settings, providing a more comprehensive understanding of the capabilities and limitations of LLMs in following attention instructions and mitigating position bias in both 3-document and 9-document settings.

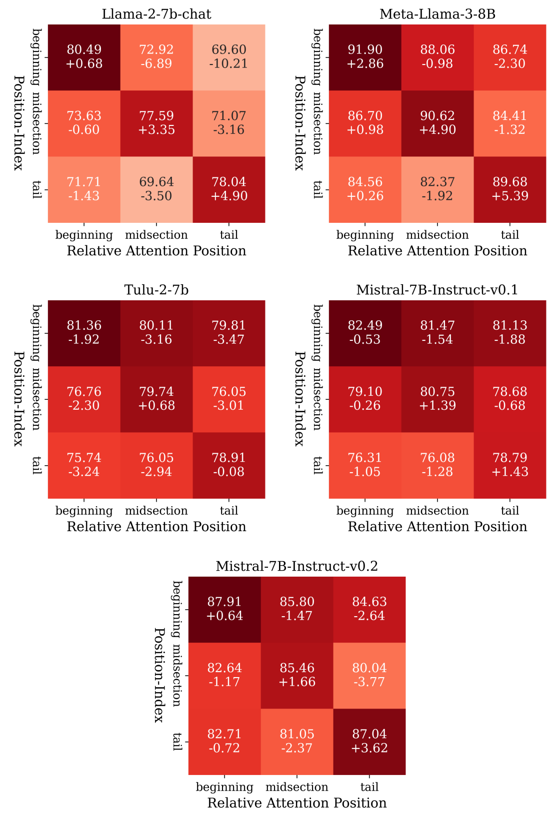

Figure 13: 3-document: absolute attention instruction under Position-index setting.

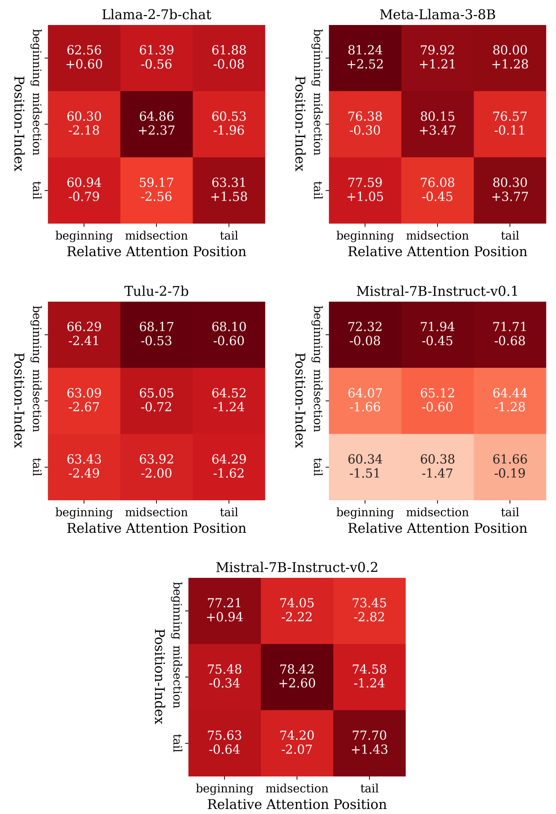

Figure 14: 9-document: absolute attention instruction under Position-index setting.