Found in the middle: Calibrating Positional Attention Bias Improves Long Context Utilization

Anonymous ACL submission

Abstract

Large language models (LLMs), even when specifically trained to process long input contexts, struggle to capture relevant information located in the middle of their input. This phenomenon has been known as the *lost-in-the-middle* problem. In this work, we make three contributions. First, we set out to understand the factors that cause this phenomenon. In doing so, we establish a connection between lost-in-the-middle to LLMs' intrinsic attention bias: LLMs exhibit an *U-shaped attention bias* where the tokens at the beginning and at the end of its input receive higher attention, regardless of their relevance. Second, we mitigate this positional bias through a calibration mechanism, *found-in-the-middle*, that allows the model to attend to contexts faithfully according to their relevance, even though when they are in the middle. Third, we show *found-in-the-middle* not only achieves better performance in locating relevant information within a long context, but also eventually leads to improved retrieval-augmented generation (RAG) performance across various tasks, outperforming existing methods by up to 10 percentage point. These findings open up future directions in understanding LLM attention bias and its potential consequences.

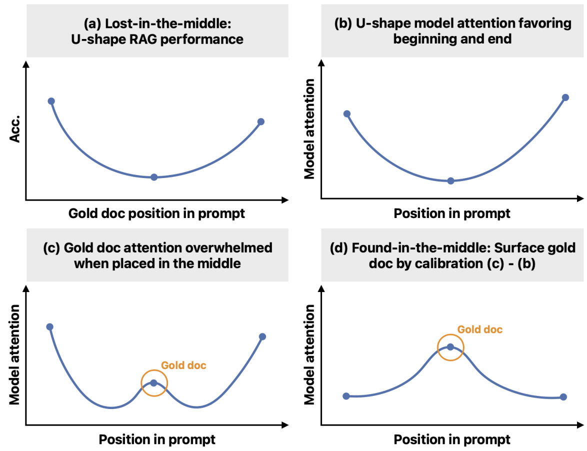

Figure 1: (a) Lost-in-the-middle refers to models' U-shape RAG performance as the relevant context's (e.g., a gold document containing the answer to a query) position varies within the input; (b) We observe models exhibit U-shape attention weights favoring leading and ending contexts, regardless of their actual contents; (c) Models do attend to relevant contexts even when placed in the middle, but are eventually distracted by leading/ending contexts; (d) We propose a calibration mechanism, find-in-the-middle, that disentangles the effect of U-shape attention bias that allows models to attend to relevant context regardless their positions.

1 Introduction

Effective prompting of large language models (LLMs) (Brown et al., 2020; Anil et al., 2023; Touvron et al., 2023) has enabled a variety of user-facing applications, including conversational interfaces (chatbots) (Thoppilan et al., 2022), search and summarization (Min et al., 2024), open-domain question answering (Izacard and Grave, 2021), tool usage (Hsieh et al., 2023), fact checking (Asai et al., 2023), and collaborative writing (Lee et al., 2019). Some of these applications, such as search and summarization (Ji et al., 2023; Min et al., 2023; Asai et al., 2023), require the ability to retrieve information from external knowledge sources. As a result, retrieval-augmented generation (RAG) has become a powerful solution. RAG fetches relevant documents (e.g. structured tables (Wang et al., 2024) and API documentation (Karpukhin et al., 2020)) from external knowledge sources and makes them available in the LLMs' input prompt (Khandelwal et al., 2020; Borgeaud et al., 2022; Izacard et al., 2022b; Xu et al., 2023a). Despite the widespread utility of RAG (Li et al., 2023a; Xiong et al., 2023; OpenAI, 2022; Gemini Team, 2023), recent experiments highlight a striking deficiency: LLMs struggle to locate relevant documents when they are placed in the middle of their input prompts (Liu et al., 2023; Li et al., 2023a). They call this the *lost-in-the-middle* phenomenon.

To overcome this phenomenon, a few mechanistic strategies have been proposed (Jiang et al., 2023; Peysakhovich and Lerer, 2023). These methods *re-rank* the relevance of different documents and *re-order* the most relevant ones to either the beginning or end of the input context. Unfortunately, re-ranking usually requires additional supervision or dedicated finetuning for performant RAG performance (Karpukhin et al., 2020; Shi et al., 2023c; Sun et al., 2023). Worse, re-ranking methods do not fundamentally improve LLMs' ability to utilize and capture relevant information from the provided input contexts. The underlying causes of this behavior remains unclear, even though it has been observed across multiple decoder-only LLMs (Touvron et al., 2023; Li et al., 2023a; OpenAI, 2022).

In this work, we make three contributions: First, we set out to understand the potential factors leading to the *lost-in-the-middle* problem. **We establish a connection between lost-in-the-middle to LLMs' intrinsic attention bias** (see Figure 1). Specifically, we find that models often demonstrate a *U-shaped* attention distributions, with higher attention values assigned to the beginning and end of the input prompt. This correlates well with the U-shaped RAG performance observed in prior literature (Liu et al., 2023). Interestingly, this focus on the beginning and end also extends to content utilization: models preferentially use information from the beginning and end of their prompts (Ravaut et al., 2023; Peysakhovich and Lerer, 2023). This leads us to hypothesize that the positional attention bias may contribute to the phenomenon, wherein the bias could lead to overreliance on content at the beginning/end of the input, regardless of its true relevance.

Second, we verify our hypothesis by intervening on this attention bias to determine its impact on performance. **We propose a mechanism to disentangle positional bias from model's attention.** We first esitmate this bias through measuring the change in attention as we vary the relative position of a fixed context in the LLM's prompt. By quantifying and then removing this bias from the attention scores for a given query, we can obtain the *calibrated attention* scores across the retrieved documents. This calibrated attention proves to be better correlated to the ground truth relevance of the document to a user query. In open-domain question answering tasks (Kwiatkowski et al., 2019), our proposed calibrated attention outperforms popular existing approaches for ranking the relevance of retrieved documents (up to 0.44 Recall@3 points). This finding challenges the recent belief that LLMs struggle to capture relevant context embedded in the middle of inputs, suggesting they may indeed be capable of doing so, but are only hindered by the overwhelming positional bias.

Third, we operationalize our calibration mechanism as a solution for this phenomenon, naming our attention intervention *found-in-the-middle*. **We show that calibrating the attention leads to improvements across two popular LLMs with different context window lengths on two RAG tasks.** Our experiments demonstrate improvements over standard model generation by up to 10 percentage point on NaturalQuestion dataset (Kwiatkowski et al., 2019). We hope the work opens up future directions in understanding LLM's attention biases and their effect on downstream tasks.

2 Positional attention bias overpowers mid-sequence context

Recent work has produced language models capable of handling increasingly long input contexts (Xiong et al., 2023; Li et al., 2023a). However, many of these models struggle to locate relevant information placed in the middle of the input sequence (Liu et al., 2023), a phenomenon known as the "lost-in-the-middle" problem. While this problem is widely recognized, the potential factors contributing to this behavior remain poorly understood. In this work, we seek to deepen our understanding of the problem through a suite of exploratory qualitative and quantitative studies.

Setup. We adhere to the original experimental setup outlined in Liu et al. (2023), utilizing an open-domain question answering task (Kwiatkowski et al., 2019) for our exploratory study. In the lost-in-the-middle setup (Liu et al., 2023), a model is tasked to answer a user query x^{q} using a set of k related documents retrieved from an external data source $D = \{x^{\mathrm{gold}}, x_1^{\mathrm{distract}}, \ldots, x_{k-1}^{\mathrm{distract}}\}$, where only the gold document x^{gold} contains the correct answer. The question and documents are typically serialized as an input sequence $x^{\mathrm{prompt}} = [x^{\mathrm{q}}, x_1^{\mathrm{doc}}, ..., x_k^{\mathrm{doc}}, x^{\mathrm{q}}]$, prompting a language model to generate the final answer[1]. Obser-

[1]We repeat the question before and after the documents so that the model can better attend to relevant contexts (Liu et al., 2023; Xu et al., 2023b).

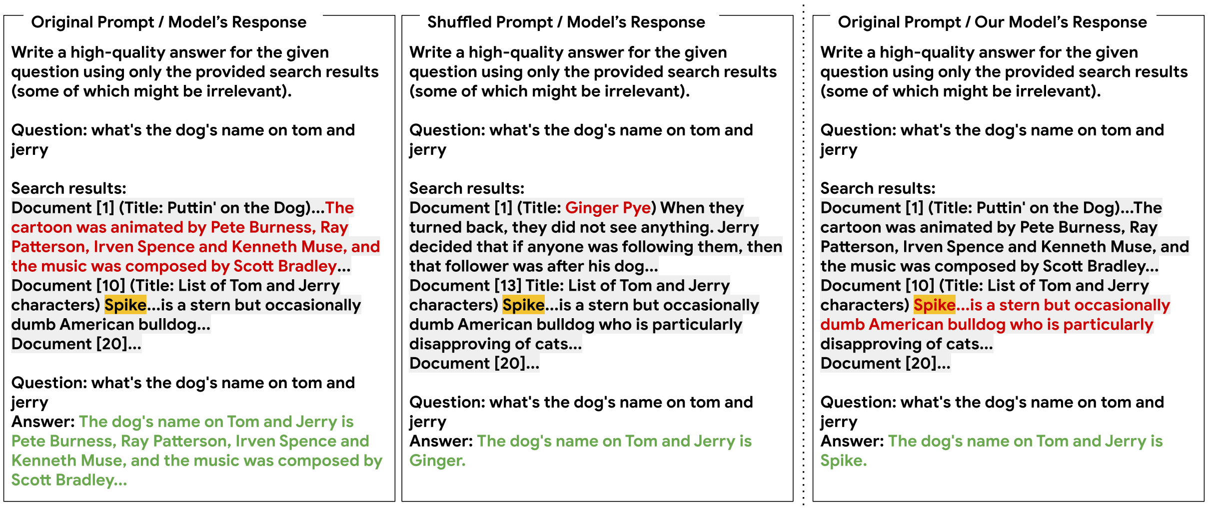

Figure 2: **Left and Middle: Qualitatively, the model's response exhibits a strong bias towards the document at the first position (red).** This persists whether the input documents retain their original order (left: gold document at the 10th position) or are randomly shuffled (middle: gold document at the 13th position). Model responses are shown in green, with the gold answer highlighted in yellow. **Right: Our attention calibration method enables the model to find relevant context even when placed in the middle.**

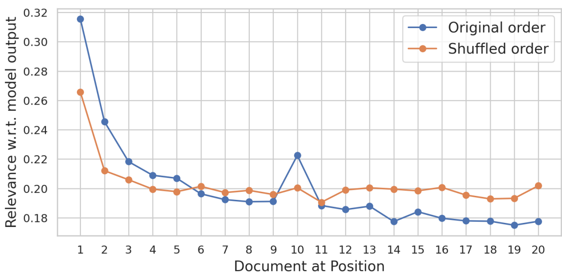

Figure 3: **Quantitatively, the model's response strongly depends on the document at the first position.** This dependence persists even after randomly shuffling the document order, irrespective of its relevance to the query. We measure this dependence by computing the TF-IDF similarity score between the response and each document (gold document originally at position 10).

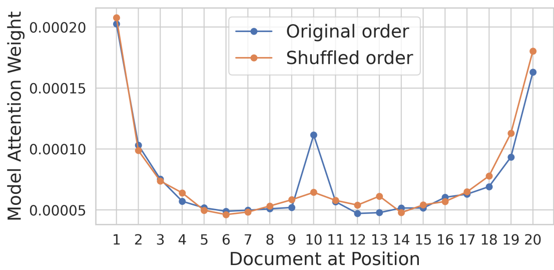

Figure 4: **Average attention weights reveal a U-shaped positional bias in the model.** Documents at the beginning and end receive greater attention, regardless of order (gold document originally at position 10). Attention is averaged across different decoder layers and attention heads.

vations indicate that model performance significantly decreases when x^{gold} is placed within the middle of the input prompt (i.e., $x^{\text{doc}}_{\lfloor k/2 \rfloor}$), compared to scenarios where x^{gold} is placed at the beginning or end. Here, we reproduce lost-in-the-middle phenomenon with a LLaMA-2-7B-chat model (Touvron et al., 2023) to gain deeper insights into the characteristics of the model's errors. We focus our error analysis on the setting where we have a total of 20 documents ($K = 20$). We specifically look at the examples where the model makes incorrect predictions when the gold document is placed at the middle (10-th) position.

2.1 U-shaped attention bias

We first examine responses generated when gold documents are placed in the **middle** of input prompts. Qualitatively, the model's response exhibits a strong bias towards the document at the first position, regardless of the gold document's location (Figure 2). This bias persists whether the input documents retain their original order or are randomly shuffled.

The strong correlation between the model's output and the first document could suggest that they are highly relevant, distracting the model (Shi et al., 2023a). However, quantitatively, the model's response strongly depends on the document at the first position (Figure 3). This dependence persists even after randomly shuffling the document order, irrespective of its relevance to the query. We measure the dependence by computing the TF-IDF similarity between the response and each document (gold document originally at position 10).

To investigate the potential origins of positional bias, we visualize the model's self-attention weights, as the weights has been shown to correlate with models' generations, although not necessarily causal (Dong et al., 2021; Zhang et al., 2023). More formally, given an input prompt consisting of K documents $x^{\text{prompt}} = [x_1^{\text{doc}}, ..., x_K^{\text{doc}}]$, where each document $x_k^{\text{doc}} = \{x_{k,i}^{\text{doc}}\}_{i=1}^{N_k}$ contains N_k tokens, let $\text{Attn} : \mathcal{X} \times \mathbb{N} \rightarrow \mathbb{R}$ denote a function that computes the average attention weights assigned to document x_k^{doc} as $\text{Attn}(x^{\text{prompt}}, k) = \sum_{i=1}^{N_k} \text{attn}(x_{k,i}^{\text{doc}})/N_k$, where $\text{attn}(x_{k,i}^{\text{doc}})$ is the attention weight value allocated to token $x_{k,i}^{\text{doc}}$ when predicting the next $|x^{\text{prompt}}| + 1$ token.

Specifically, we visualize the average self-attention weights assigned to each document across all tokens, decoder layers, and heads. We investigate how these weights vary based on document position within the input prompt. Interestingly, Figure 4 (blue curve) reveals a U-shaped attention pattern. Documents near the beginning and end of the input receive higher weights, while those in the middle receive lower weights. Crucially, the U-shaped pattern persists even after randomly shuffling document order (Figure 4, orange curve), suggesting that this bias does not depend on the documents' actual content.

2.2 Does attention favor relevant context?

Observation 1: Model prioritizes relevant contexts from the same position. In Figure 4, we observe a significant difference in attention values at x_{10}^{doc} when comparing examples with original document order (blue) and randomly shuffled order (orange). Specifically, the attention value is notably higher when when x_{10}^{doc} is controlled to be x^{gold}. This contrasts with instances where x_{10}^{doc} is uncontrolled, suggesting that apart from U-shaped positional bias, the model exhibits an ability to *prioritize* relevant context.

Observation 2: Model prioritizes highly-weighted documents for generation. Based on these observations, we hypothesize that positional attention bias significantly influence the model's tendency to rely heavily on the first documents during output generation. Specifically, the models are more likely to incorporate the document receiving the highest attention (often the first) into its output. To validate this, for each of the examples of interest, we divide their documents into first half receiving higher model attention and second half receiving lower attention. We then count the number of examples in which the first or second half contains the document that is most likely used in the model's generation (i.e., having the highest TF-IDF score with model's response). In Table 1, we show that documents receiving higher attention positively correlates with them being used in the model's generation.

Table 1: Number of examples where the most likely used document in the model's generation falls within the first half of documents receiving higher model attention or second half receiving lower attention. We see that there is a strong correlation where documents receiving higher attention are more likely to be used in model's response.

	Most Likely Used	
	# of examples	%
Highest Half Attention	490	71%
Lowest Half Attention	200	29%

From the above studies, we see that not only the model exhibits a U-shape positional attention bias, but this bias also correlates strongly with the model's biased tendency in using documents placed at certain positions in forming its response. We thus conjecture that lost-in-the-middle happens because of the dominating force of positional bias.

3 Find-in-the-middle: modeling and isolating positional attention bias

Ideally, a model should leverage contexts in the input prompts—faithfully according to their relevance—for generating the response, instead of biasing towards contexts placed at certain positions within the input. Towards this goal, we are interested in modeling the positional attention bias and mitigating it such that model attention can reflect the true relevance of the input context and ultimately improve models' effective utilization of the full context window.

3.1 Two main factors in model attention

In Sec. 2, we find that there are two main forces driving the model attention assigned to different documents of an input prompt: (a) where the document locates within the entire input, and (b) the relevance of the document.

Our hypothesis. We thus consider modeling the observable attention weights allocated to the k-th document of an input x^{prompt} as:

$$\text{Attn}(x^{\text{prompt}}, k) = f(\text{rel}(x_k^{\text{doc}}), \text{bias}(k)), \quad (1)$$

Table 2: High correlations between model attention with document relevance and positional bias supports our hypothesized model.

Hypothesis test	$\mathrm{rel}(x^{\mathrm{doc}})$	$\mathrm{bias}(k)$	% of valid pairs
Condition 1	Fixed	Varying	82%
Condition 2	Varying	Fixed	75%

where $\mathrm{rel}(\cdot)$ measures the relevance of an input document, $\mathrm{bias}(\cdot)$ characterizes the positional attention bias, $f(\cdot)$ is some unknown monotonically increasing function w.r.t. to both $\mathrm{rel}(x_k^{\mathrm{doc}})$ and $\mathrm{bias}(k)$. For ease of exposition, in the remainder of the paper, we overload $\mathrm{Attn}(x^{\mathrm{doc}}, k)$ to denote the attention value assigned to document x_k^{doc} placed at the k-th position within an input prompt containing K documents.

Corroborating our assumed model. Here, we conduct a suite of controlled experiments using NaturalQuestion with $K = 20$ and a LLaMA-2-7B-chat model to corroborate our assumed model. Specifically, for Eq. 1 to hold, it implies that:

Condition 1: When the relevance term is fixed, model attention increases as positional bias increases. That is, given two documents x^{doc1} and x^{doc2}: *if* $\mathrm{Attn}(x^{\mathrm{doc1}}, k) > \mathrm{Attn}(x^{\mathrm{doc1}}, l)$, *then* $\mathrm{Attn}(x^{\mathrm{doc2}}, k) > \mathrm{Attn}(x^{\mathrm{doc2}}, l)$.

Condition 2: Similarly, when the document position k is fixed, model attention increases as the relevance of the document increase: *if* $\mathrm{Attn}(x^{\mathrm{doc1}}, k) > \mathrm{Attn}(x^{\mathrm{doc2}}, k)$, *then* $\mathrm{Attn}(x^{\mathrm{doc1}}, l) > \mathrm{Attn}(x^{\mathrm{doc2}}, l)$.

We validate Condition 1 and 2 on 100 randomly sampled examples from NaturalQuestion dataset, each with $K = 20$ documents. For validating Condition 1, given a pair of documents $(x^{\mathrm{doc1}}, x^{\mathrm{doc2}})$ and positions (k, l), we can compute whether the relationship holds across all possible pairs. We can similarly test for Condition 2. In Table 2, we see that the percentage of valid example pairs are decently high, 82% and 75% respectively, for both conditions, providing supports to our hypothesis.

Recall that our goal is to disentangle positional attention bias from model attention such that the model can faithfully attend to relevant contexts, independent from their positions. So far, while we have established the monotonic increasing nature of f in Eq. 1, we have yet characterize the actual form of f to remove the positional bias term from model attention.

To approximate f, we consider simple linear models by following machine learning principles (a.k.a. Occam's razor), for robust estimation:

$$\mathrm{Attn}(x^{\mathrm{doc}}, k) = \mathrm{rel}(x^{\mathrm{doc}}) + \mathrm{bias}(k) + \epsilon, \quad (2)$$

where ϵ is a noise.

To test how the model captures the underlying relationship, we compute Spearman's rank correlation between $\mathrm{Attn}(x^{\mathrm{doc1}}, k) - \mathrm{Attn}(x^{\mathrm{doc2}}, k)$ and $\mathrm{Attn}(x^{\mathrm{doc1}}, l) - \mathrm{Attn}(x^{\mathrm{doc2}}, l))$ over quadruplets of $(x^{\mathrm{doc1}}, x^{\mathrm{doc2}}, k, l)$ collected from NaturalQuestion. A high correlation indicates small discrepancy between $\mathrm{Attn}(x^{\mathrm{doc1}}, k) - \mathrm{Attn}(x^{\mathrm{doc2}}, k)$ and $\mathrm{Attn}(x^{\mathrm{doc1}}, l) - \mathrm{Attn}(x^{\mathrm{doc2}}, l))$. From our study, the linear model results in decently high correlation, 0.763, suggesting its effectiveness despite the simplicity. We therefore adopt Eq. 2 as our model and leave other alternatives with more degree of freedoms as future work [2].

3.2 Disentangling positional attention bias

Most notably, having a simple form of f allows us to isolate the effect of positional bias from model attention. Specifically, following from Eq. 2, we can first obtain a reference model attention value with a dummy document x^{dum} by:

$$\mathrm{Attn}(x^{\mathrm{dum}}, k) = \mathrm{rel}(x^{\mathrm{dum}}) + \mathrm{bias}(k) + \epsilon. \quad (3)$$

By subtracting Eq. 2 and Eq. 3, we can offset the bias term and obtain:

$$\begin{aligned} &\mathrm{rel}(x^{\mathrm{doc}}) \quad (4) \\ &= \mathrm{Attn}(x^{\mathrm{doc}}, k) - \mathrm{Attn}(x^{\mathrm{dum}}, k) + \mathrm{rel}(x^{\mathrm{dum}}) \end{aligned}$$

Consider using a consistent dummy document x^{dum} which has a constant $\mathrm{rel}(x^{\mathrm{dum}})$, we are then able to obtain the true relevance of different documents x^{doc}, free from the positional bias. We refer to $\mathrm{Attn}(x^{\mathrm{doc}}, k) - \mathrm{Attn}(x^{\mathrm{dum}}, k)$ as *calibrated attention* as it removes the baseline attention.

Calibrated attention finds relevant contexts in the middle. Eq. 4 allows us to leverage calibrated attention to estimate and rank the relevance of different documents within an input prompt. To validate the effectiveness of our model, we evaluate using calibrated attention to re-rank documents in an input prompt w.r.t. a given query. We evaluate on NaturalQuestion where we focus on the most challenging setting when the gold document in placed in the middle of the input prompt. We compare our model to:

[2] In Appendix C, we also explore log-linear models, which results in competitive 0.76 rank correlation.

Table 3: Calibrated attention outperforms existing methods in ranking the relevance of retrieved contexts given a user query. We report Recall@3 on NaturalQuestion when gold documents are placed in the middle of input context.

	Number of total documents	
Method	$K = 10$	$K = 20$
Vanilla attention	0.3917	0.1340
Query generation	0.6474	0.5378
Relevance generation	0.5152	0.3578
Calibrated attention	**0.7163**	**0.5706**

- Vanilla attention: Using uncalibrated attention $\mathrm{Attn}(x^{\mathrm{prompt}}, k)$ to rank the documents.
- Query generation (Sun et al., 2023): Using likelihood of the model in generating the query based on the document.
- Relevance generation (Sun et al., 2023): Prompting the model to answer whether a document is relevant to a query.

In Table 3, we compare Recall@3 of different methods where we vary the total number of documents retrieved. We see that the proposed calibrated attention consistently outperforms vanilla attention by a large margin, and also shows superior performances when compared to the other two re-ranking metrics. The results validate that our proposed modeling approach is effective, and that if calibrated appropriately, language models can locate relevant information even when they are hidden in the middle of the input.

4 Improving long-context utilization with attention calibration

Having validated that calibrated attention through find-in-the-middle is effective in locating relevant information within a long input context, we are ultimately interested in leveraging it to tackle lost-in-the-middle problem and practically improve a model's RAG performance.

4.1 Attention calibration

To allow the model to attend to contexts without being dictated by positional bias, we propose to intervene the model's attention based on the proposed calibrated attention. Specifically, given an input x^{prompt}, instead of allocating $\mathrm{rel}(x_k^{\mathrm{doc}}) + \mathrm{bias}(k)$ attention to the k-th document, our ideal model attention $\mathrm{Attn}_{\mathrm{calibrated}}(x_k^{\mathrm{doc}})$ would reflect only the relevance of the context $\mathrm{rel}(x_k^{\mathrm{doc}})$.

To achieve this, we propose to redistribute the attention values assigned to $\{x_k^{\mathrm{doc}}\}_{k=1}^{K}$ according to $\mathrm{rel}(x_k^{\mathrm{doc}})$. Specifically, for each document x_k^{doc}, we propose to rescale the attention values on the tokens within the document, $\{x_{k,i}^{\mathrm{doc}}\}_{i=1}^{N_k}$, by:

$$\mathrm{attn}_{\mathrm{calibrated}}(x_{k,i}^{\mathrm{doc}}) = \frac{\alpha_k}{\mathrm{Attn}_{\mathrm{original}}(x_k^{\mathrm{doc}})} \cdot \mathrm{attn}_{\mathrm{original}}(x_{k,i}^{\mathrm{doc}}) \cdot C, \tag{5}$$

where $\alpha_k = \mathrm{Softmax}(\mathrm{rel}(x_k^{\mathrm{doc}}), t)$, t is the temperature hyperparamter, and C is a normalization constant to ensure the total attention $\sum_{k,i} x_{k,i}^{\mathrm{doc}}$ remains unchanged. With the rescaling, we effectively make the final attention on x_k^{doc}:

$$\mathrm{Attn}_{\mathrm{calibrated}}(x_k^{\mathrm{doc}}) \propto \mathrm{Softmax}(\mathrm{rel}(x_k^{\mathrm{doc}}), t), \tag{6}$$

where higher attention is allocated to more relevant context, and t controls the disparity level.

4.2 Calibrated v.s. uncalibrated attention

We evaluate the performance of the proposed attention calibration method. We conduct experiments on two multi-document question answering tasks (more details in Appendix A), NaturalQuestion (Kwiatkowski et al., 2019) and SynthWiki (Peysakhovich and Lerer, 2023), with two models supporting different context window length: `LLaMA-2-7B-chat` (LLaMA) (Touvron et al., 2023) and `Vicuna-7b-v1.5-16k` (Vicuna) (Li et al., 2023a) with 4k and 16k context window respectively. For each dataset, we consider two settings with different number of retrieved documents, $K = \{10, 20\}$. We leave further implementation details in Appendix B.

Attention calibration improves long-context utilization across various datasets and models. In Figure 5, we see that our proposed calibrated-attention intervention method consistently outperforms the uncalibrated baseline by a large margin (up to 10 percentage point (pp) improvement) across different tasks and models. On the most challenging scenario when the gold document is placed mid-sequence, attention calibration consistently offers improvements from 6-10 pp. Notably, we see that attention calibration's performance curve lies entirely above the vanilla baseline curve, validating the effectiveness of our method in improving models' long context utilization.

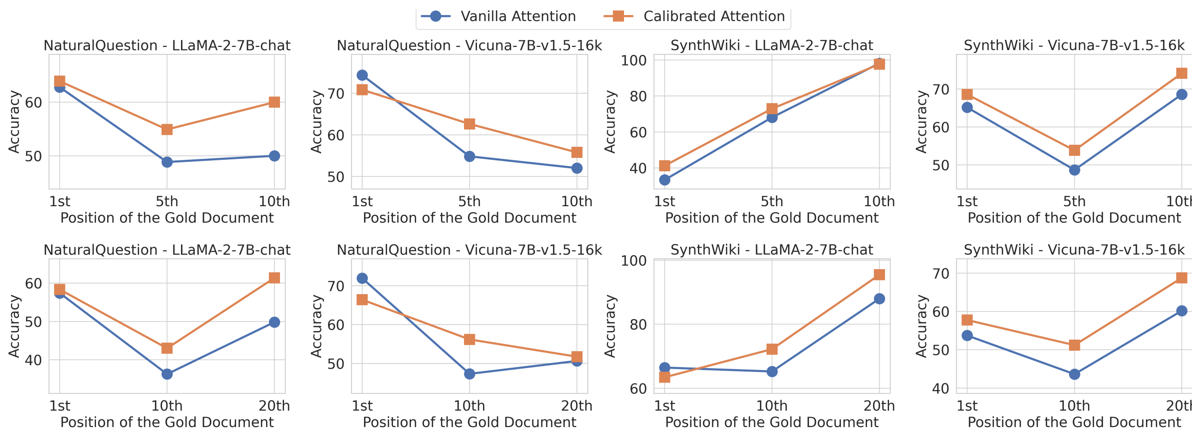

Figure 5: **Attention calibration effectively improves models' context utilization ability, with its performance curves lying above standard vanilla attention.** Top/Bottom row: 10/20-doc. Numbers shown in Table 5.

4.3 Attention calibration in practice

In practice, to avoid the lost-in-the-middle effect, one commonly adopted workaround is to reorder the document positions, where documents considered more relevant are placed towards the beginning (or end) of the input. While these methods have led to performance improvements over the baseline without reordering, without handling the model's intrinsic bias, reordering-based methods' performance relies heavily on the correct ranking of the documents. We are thus interested in validating whether attention calibration can be applied on top of re-ordering methods to provide another layer of improvements.

Attention calibration improves existing RAG pipelines. We continue using NaturalQuestion and SynthWiki for evaluation. We compare to existing reordering methods including:

- Prompt reordering (Sun et al., 2023; Liang et al., 2023): Reorder documents based on relevance score generated through prompting.
- LongLLMLingua-r_k (Jiang et al., 2023): Reorder documents using query generation as the reranking metric.
- Attention sorting (Peysakhovich and Lerer, 2023): Reorder documents using vanilla model attention assigned to the documents.

In Figure 6, we note that LongLLMLingua-r_k and prompt reodering are invariant to the gold document's position since they compute the relevance of each document independently. First, we see that reordering methods do alleviate lost-in-the-middle problem where models' performances increase when gold documents is placed mid-sequence. Furthermore, by applying attention calibration on top of a reodring mechanism (LongLLMLingua-$r_k$in this case), LongLLMLingua-r_k with calibration consistently achieve the highest performance across datasets and models, suggesting a way to further improve current RAG pipeline.

5 Related work

Retrieval augmented generation. While LLMs exhibit strong capabilities (Gemini Team, 2023; OpenAI, 2022), their knowledge is inherently limited in its pretraining data, and they are observed to struggle in handling knowledge intensive tasks (Petroni et al., 2020). To tackle this, retrieval augmented generation (RAG) is an effective framework that retrieves relevant information from external knowledge sources to aid and ground language models' generation (Lewis et al., 2020; Khandelwal et al., 2020; Borgeaud et al., 2021; Izacard and Grave, 2021; Izacard et al., 2022b).

Although RAG has powered many recent language model applications from question-answering (Izacard and Grave, 2021) to automatic task completion (Shen et al., 2023), recent work show that LLMs tend to *lost-in-the-middle*, significantly hindering the full potential of RAG (Liu et al., 2023). In this work, we take a step further to understand the lost-in-the-middle problem from the viewpoint of attention bias. Moreover, we propose a remedy through attention calibration, which improves upon existing RAG frameworks.

Long-context utilization in language models. There is a rich literature on enabling LLMs to handle longer input contexts, including designing efficient training and finetuning schemes (Dao et al.,

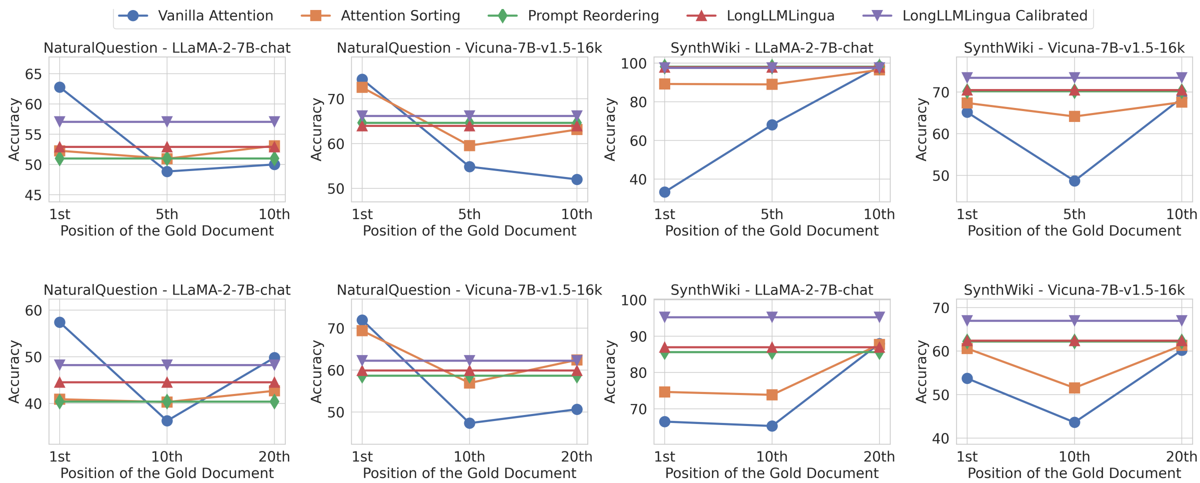

Figure 6: **Attention calibration can be applied on top of reordering-based methods to provide further performance boost.** Top/Bottom row: 10/20-doc. Numbers shown in Table 5.

2022; Li et al., 2023b,a; Shi et al., 2023b) and inference-time methods that extend an LLM's context length (Press et al., 2021; Ratner et al., 2023; Xiao et al., 2023; Bertsch et al., 2023). Nonetheless, even models specifically trained for long-context suffer lost-in-the-middle problem (Liu et al., 2023; Li et al., 2023a).

To improve LLMs' performance on handling long contexts, recent methods design better prompting techniques and pipelines that mechanically work around the lost-in-the-middle problem (Chen et al., 2023; Jiang et al., 2023; Peysakhovich and Lerer, 2023; Junqing et al., 2023). For instance, to avoid having the models process long input contexts, (Chen et al., 2023; Junqing et al., 2023) proposes to split long inputs into shorter contexts for models to better understand. To avoid relevant context being missed by the model, (Jiang et al., 2023; Peysakhovich and Lerer, 2023) proposes to rank the relevance of different parts of the input and re-order the most important parts to either the beginning or end of the entire input, where the models tend to focus more.

While these existing solutions lead to improved model performances by manipulating the input contexts, they do not fundamentally improve LLMs' underlying long-context utilization capability. In contrast, we set out to directly improve LLMs' long-context utilization capability to mitigate lost-in-the-middle problem.

Self-attention and attention bias. The attention mechanism is initially introduced in RNN-based encoder-decoder architectures (Bahdanau et al., 2015; Luong et al., 2015). Building upon the self-attention mechanism, transformers (Vaswani et al., 2017) have achieved state-of-the-art performance in various domains (Devlin et al., 2018; Dosovitskiy et al., 2020). Self-attention has also been widely used as a proxy to understand and explain model behaviors (Clark et al., 2019; Hao et al., 2021; Vashishth et al., 2019).

However, the relationship between the lost-in-the-middle problem and LLM's self-attention has been under-explored. As an initial trial, "attention sorting" (Peysakhovich and Lerer, 2023) sorts documents multiple times by the attention they receive to counter lost-in-the-middle. Recently, He et al. (2023) construct a dataset for training LLMs to focus on the most relevant documents among long contexts. Unlike the method, which necessitate significant investment in data collection and LLM tuning, our method offers an efficient solution by mitigating lost-in-the-middle problem with off-the-shelf LLMs.

6 Discussion

In this work, we understand and address the lost-in-the-middle phenomenon, by establishing a connection between the phenomenon and models' positional attention bias. We mitigate the bias by attention calibration which directly modifies the model's attention mechanism, enabling LLMs to more faithfully attend to contexts based on their relevance, rather than their position. Experiments show that attention calibration improves the performance compared to its uncalibrated counterpart especially when relevant context occurs in the middle of the input. We additionally show attention calibration can be applied on top of existing reordering pipelines to further improve models' performance.

Limitations

While our study presents significant advances in addressing the "lost-in-the-middle" problem and improving RAG performance in LLMs, several limitations are noteworthy:

Simplification of the mechanism behind positional attention bias. We proposed a simple hypothesis to model the positional attention bias, as shown in Eq. 1. However, the intrinsic mechanisms that drive this bias could be more intricate and dynamic than our current model accounts for. It is possible that some aspects of attention bias are learnable or adaptive, responding to subtle aspects of the data or training process that our current approach does not consider.

Computational overhead. Our method of calibrating positional attention bias, while effective, introduces additional computational overhead. Specifically, we require extra $O(K)$ model forward passes to calibrate attention at each position, compared to vanilla model generation. However, in this study we aim to discover and calibrate the positional attention bias from a scientific perspective. We expect that our discovery can enable future research into developing more calibration methods with lower computational overhead.

Positional attention bias may be beneficial. Our method aims to completely remove positional attention bias. However, it is important to note that this positional bias might actually be beneficial in certain contexts. In some specific tasks or scenarios, the natural tendency of models to focus more on the beginning and end of inputs could align well with the structure of the task or the nature of the data. Therefore, understanding the tasks and the applications is required before adopting our proposed calibration method.

The root cause of attention bias is unclear. In this work, we aim to discover and understand the connection between the lost-in-the-middle problem and LLMs' intrinsic attention bias. However, our work does not definitively pinpoint the root cause of attention bias in LLMs. The cause of such a bias could be attributed to the distribution of pretraining corpora, the transformer model architecture, and the optimization process. Future research needs to delve deeper into the origins of this phenomenon.

Ethical Statement

In our research, we focus on enhancing the performance of large language models using existing public datasets, ensuring that no personal or sensitive data was collected or utilized. Our attention calibration method is aimed at improving the efficiency and accuracy of retrieval-augmented generation, with potential benefits across various domains including search engines, question-answering systems, and other text-based applications. It is important to acknowledge that as our technique builds upon pre-trained language models, it may inadvertently inherit and propagate existing biases inherent in these models. Apart from this significant concern, we do not identify any other immediate risks arising from the methodologies or findings presented in our paper.

References

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Amanda Bertsch, Uri Alon, Graham Neubig, and Matthew R Gormley. 2023. Unlimiformer: Long-range transformers with unlimited length input. *arXiv preprint arXiv:2305.01625*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2021. Improving language models by retrieving from trillions of tokens. *arXiv preprint arXiv:2112.04426*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Howard Chen, Ramakanth Pasunuru, Jason Weston, and Asli Celikyilmaz. 2023. Walking down the memory maze: Beyond context limit through interactive reading. *arXiv preprint arXiv:2310.05029*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yue Dong, Chandra Bhagavatula, Ximing Lu, Jena D Hwang, Antoine Bosselut, Jackie Chi Kit Cheung, and Yejin Choi. 2021. On-the-fly attention modulation for neural generation. *arXiv preprint arXiv:2101.00371*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gemini Team. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2021. Self-attention attribution: Interpreting information interactions inside transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12963–12971.

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 e-prints*, pages arXiv–2311.

Cheng-Yu Hsieh, Si-An Chen, Chun-Liang Li, Yasuhisa Fujii, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. 2023. Tool documentation enables zero-shot tool-usage with large language models. *arXiv preprint arXiv:2308.00675*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022a. 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. Association for Computational Linguistics.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022b. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

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

Huiqiang Jiang, Qianhui Wu, , Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu.

2023. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. *ArXiv preprint*, abs/2310.06839.

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

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

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

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300*.

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.

Dacheng Li, Rulin Shao, Anze Xie, Ying Sheng, Lianmin Zheng, Joseph E. Gonzalez, Ion Stoica, Xuezhe Ma, and Hao Zhang. 2023a. How long can opensource llms truly promise on context length?

Dacheng Li, Rulin Shao, Anze Xie, Eric P Xing, Joseph E Gonzalez, Ion Stoica, Xuezhe Ma, and Hao Zhang. 2023b. Lightseq: Sequence level parallelism for distributed training of long context transformers. *arXiv preprint arXiv:2310.03294*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Sewon Min, Suchin Gururangan, Eric Wallace, Hannaneh Hajishirzi, Noah A. Smith, and Luke Zettlemoyer. 2024. SILO language models: Isolating legal risk in a nonparametric datastore. In *ICLR*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *EMNLP*.

OpenAI. 2022. Chatgpt.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. 2020. Kilt: a benchmark for knowledge intensive language tasks. *arXiv preprint arXiv:2009.02252*.

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

Ofir Press, Noah Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*.

Nir Ratner, Yoav Levine, Yonatan Belinkov, Ori Ram, Inbal Magar, Omri Abend, Ehud Karpas, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. Parallel context windows for large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6383–6402, Toronto, Canada. Association for Computational Linguistics.

Mathieu Ravaut, Shafiq Joty, Aixin Sun, and Nancy F Chen. 2023. On position bias in summarization with large language models. *arXiv preprint arXiv:2310.10570*.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. In *Advances in Neural Information Processing Systems*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023a. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Victoria Lin, Noah A Smith, Luke Zettlemoyer, Scott Yih, and Mike Lewis. 2023b. In-context pretraining: Language modeling beyond document boundaries. *arXiv preprint arXiv:2310.10638*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023c. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937, Singapore. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

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 arXiv:2307.09288*.

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention interpretability across nlp tasks. *arXiv preprint arXiv:1909.11218*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, and Tomas Pfister. 2024. Chain-of-table: Evolving tables in the reasoning chain for table understanding. In *International Conference on Learning Representations*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *arXiv*.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, et al. 2023. Effective long-context scaling of foundation models. *arXiv preprint arXiv:2309.16039*.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2023a. Retrieval meets long context large language models. *arXiv preprint arXiv:2310.03025*.

Xiaohan Xu, Chongyang Tao, Tao Shen, Can Xu, Hongbo Xu, Guodong Long, and Jian-guang Lou. 2023b. Re-reading improves reasoning in language models. *arXiv preprint arXiv:2309.06275*.

Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. 2023. Tell your model where to attend: Post-hoc attention steering for llms. *arXiv preprint arXiv:2311.02262*.

A Multi-doc QA datasets

We use NaturalQuestions (Kwiatkowski et al., 2019)[3] (released in Apache-2.0 license) and SynthWiki (Peysakhovich and Lerer, 2023)[4] to conduct the experiments. Both datasets contains question-answer pairs, a gold document contains the answer, and $K-1$ distractor documents, where $K=10$ and 20.

The NaturalQuestions dataset is the subset with 2655 queries selected by Liu et al. (2023)[5] where the annotated long answer is a paragraph. The $k-1$ distractor passages are Wikipedia chunks retrieved by Contriever (Izacard et al., 2022a) that are most relevant to the query but do not contain any of the annotated answers in NaturalQuestions. The distractor documents are presented in the context in order of decreasing relevance.

The SynthWiki dataset (Peysakhovich and Lerer, 2023) is a synthetic multi-doc QA dataset with 990 entries. All the documents in SynthWiki are GPT-4 generated Wikipedia paragraphs for fictional people, thus it can minimize the knowledge contamination issue from pre-training and ensure the LLMs can only use information from the provided context. The distractor documents are randomly sampled and randomly ordered in SynthWiki.

NaturalQuestions is collected from public English Wikipedia articles and SynthWiki is collected by GPT-4 automatic generation of English fake Wikipedia articles. These two dataset should not contain any information that names or uniquely identifies individual people or offensive content. We ensure that the use of these two datasets was consistent with their intended purpose for academic research and in accordance with their specified licensing agreements.

B Implementation details

In our experiments, we utilize `LLaMA-2-7B-chat` and `Vicuna-7b-v1.5-16k` as the base models. Both models consist of 32 decoder layers, each with 32 attention heads. In applying attention calibration method to intervene model attention, we apply only to the last 16 decoder layers (and all of their attention heads). We find that intervening early layers may lead to unstable generation. We leave finding the best set of attention heads to intervene as future directions (Zhang et al., 2023).

In the experiments, we find attention calibration to be robust to the temperature term t in Eq. 5. We set $t = 5\mathrm{e}{-5}$ for all experiments.

C Additional experiment results

Different model formulations. To approximate (1), in addition to linear models as shown in (2), we also investigate log-linear models, which is defined as

$$\log \mathrm{Attn}(x^{\mathrm{doc}}, k) = \mathrm{rel}(x^{\mathrm{doc}}) + \mathrm{bias}(k) + \epsilon, \quad (7)$$

where ϵ is a noise. We compute rank correlation as described in Sec. 3. The result is shown in Table 4. The log-linear model and linear are competitive to each other, which all result in rank correlation above 0.76.

Table 4: Rank correlations of linear and log-linear models.

Model form of f	Rank correlation
Linear	0.7633
Log-linear	0.7605

Experiment tables. Table 5 shows the exact numbers in our experiments.

D Compute and inference details

In the experiments, we use the Huggingface Transformer package[6] with the two models: LLaMA-2-7B-chat[7] and Vicuna-7B-v1.5-16k[8] both contains 7B parameters. We run the experiments with two NVIDIA A100 GPUs. The inference time is roughly 1 to 3 hours on both datasets. We run our experiments with all greedy decoding without any non-deterministic factor, so we only need to run the experiments for once. Our method is a pure inference method, so there is no need to do training or hyperparameter searching.

[3] https://github.com/google-research-datasets/natural-questions

[4] https://github.com/adamlerer/synthwiki

[5] https://github.com/nelson-liu/lost-in-the-middle

[6] https://github.com/huggingface/transformers

[7] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

[8] https://huggingface.co/lmsys/vicuna-7b-v1.5-16k

Table 5: Our proposed attention intervention by calibrated attention stably improves models' RAG performances compared to existing re-ordering based baselines.

			Gold position in 10 documents				Gold position in 20 documents			
Dataset	Model	Method	1st	5th	10th	Avg.	1st	10th	20th	Avg.
NaturalQuestion	LLaMA	Vanilla attention	62.78	48.85	50.01	53.88	57.40	36.27	49.83	47.83
		Calibrated attention	63.91	54.91	60.00	59.61	58.34	43.01	61.35	54.23
		Attention sorting	52.27	50.96	53.10	52.11	40.90	40.3	42.71	41.30
		Prompt reordering	-	-	-	51.00	-	-	-	40.41
		LongLLMLingua-r_k	-	-	-	52.92	-	-	-	44.56
		LongLLMLingua-r_k + Cal.	-	-	-	57.06	-	-	-	48.24
	Vicuna	Vanilla attention	74.35	54.83	52.01	60.39	71.93	47.34	50.65	56.64
		Calibrated attention	70.84	62.61	55.78	63.07	66.40	56.19	51.75	58.11
		Attention sorting	72.54	59.54	63.12	65.06	69.37	56.91	62.41	62.89
		Prompt reordering	-	-	-	64.63	-	-	-	58.68
		LongLLMLingua-r_k	-	-	-	63.95	-	-	-	59.92
		LongLLMLingua-r_k + Cal.	-	-	-	66.17	-	-	-	62.22
SynthWiki	LLaMA	Vanilla attention	33.33	68.08	98.18	66.53	66.46	65.25	87.97	73.22
		Calibrated attention	41.21	72.92	97.67	70.60	63.43	72.22	95.45	77.03
		Attention sorting	89.19	88.98	96.46	91.54	74.64	73.83	87.71	78.73
		Prompt reordering	-	-	-	98.18	-	-	-	85.65
		LongLLMLingua-r_k	-	-	-	97.87	-	-	-	86.96
		LongLLMLingua-r_k + Cal.	-	-	-	97.57	-	-	-	95.25
	Vicuna	Vanilla attention	65.15	48.68	68.58	60.80	53.73	43.63	60.20	52.52
		Calibrated attention	68.58	53.83	74.14	65.52	57.77	51.21	68.78	59.25
		Attention sorting	67.37	64.14	67.57	66.36	60.60	51.55	61.31	57.82
		Prompt reordering	-	-	-	70.20	-	-	-	62.22
		LongLLMLingua-r_k	-	-	-	70.50	-	-	-	62.42
		LongLLMLingua-r_k + Cal.	-	-	-	73.43	-	-	-	66.96