CAFE: Retrieval Head-based Coarse-to-Fine Information Seeking to Enhance Multi-Document QA Capability

Anonymous ACL submission

Abstract

Advancements in Large Language Models (LLMs) have extended their input context length, yet they still struggle with retrieval and reasoning in long-context inputs. Existing methods propose to utilize the prompt strategy and Retrieval-Augmented Generation (RAG) to alleviate this limitation. However, they still face challenges in balancing retrieval precision and recall, impacting their efficacy in answering questions. To address this, we introduce **CAFE**, a two-stage coarse-to-fine method to enhance multi-document question-answering capacities. By gradually eliminating the negative impacts of background and distracting documents, CAFE makes the responses more reliant on the evidence documents. Initially, a coarse-grained filtering method leverages retrieval heads to identify and rank relevant documents. Then, a fine-grained steering method guides attention to the most relevant content. Experiments across benchmarks show that CAFE outperforms baselines, achieving an average SubEM improvement of up to 22.1% and 13.7% over SFT and RAG methods, respectively, across three different models.

1 Introduction

Researchers have undertaken various efforts to extend the context length of Large Language Models (LLMs), ranging from advancements in model architectures (Yen et al., 2024; Munkhdalai et al., 2024) to optimizations in training methods (Fu et al., 2024b; An et al., 2024; Xiong et al., 2024). These developments have enabled some recently introduced LLMs to support relatively long context inputs (*i.e.,* 128K context length for LLaMA-3.1 (Dubey et al., 2024) and Qwen-2.5 (Yang et al., 2024), and even 10M context length for Gemini (Reid et al., 2024)). However, recent studies indicate that LLMs exhibit limitations in retrieval and reasoning capability when processing the long context input (Liu et al., 2024; Lee et al., 2024; Wang et al., 2024; Li et al., 2024c), which poses significant challenges for their effective application in downstream tasks, including book summarization (Bai et al., 2024a), multi-document question answering (Zhang et al., 2024c), and code repository understanding (Bai et al., 2024b).

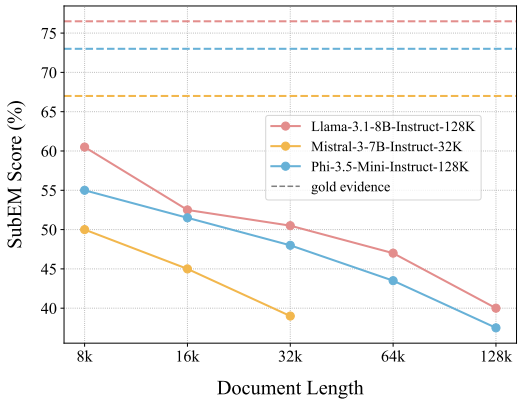

Figure 1: LLMs' performance on HotpotQA varies with the number of input documents. Solid lines represent performance with the gold document, while dashed lines show performance as more documents are added.

In long-context reasoning scenarios, particularly multi-document question answering tasks (Zhu et al., 2024), the performance of LLMs degrades significantly as context length increases, especially when compared to using only the gold evidence documents, as illustrated in Figure 1. To mitigate this issue, existing studies often adopt identify-then-reason approaches. One line of work leverages external retrieval models or prompts LLMs to extract relevant information from long inputs (Agrawal et al., 2024; Zhang et al., 2024b; Jiang et al., 2024). However, these methods are typically constrained by the limited capabilities of external retrievers or the instruction-following capacity of LLMs, often resulting in low recall. Another line of work enhances the retrieval capacity of specific attention heads (typically retrieval heads) through fine-tuning, using them to locate gold evidence within

long contexts. However, this approach often demands substantial training data and exhibits limited generalization to out-of-domain contexts.

To better tackle multi-document question answering tasks, we draw inspiration from the human problem-solving process, which typically unfolds in three phases. (1) Identification: selecting a subset of relevant documents from the entire collection to form a manageable candidate set; (2) Focusing: further selecting and paying more attention to the information most helpful for answering the question from the candidate set; and (3) Reasoning: leveraging the hierarchical information gathered in the previous two phases to perform reasoning and derive the final answer. Inspired by this, we propose a three-phase framework that follows the identification–focusing–reasoning paradigm. We begin by empirically analyzing the effectiveness of different identification methods and the information flow across various segments within attention modules. Our key observations are as follows. First, leveraging attention scores from specific retrieval heads provides a strong and effective signal for identifying relevant documents. Second, modifying the attention between the question and the evidence documents influences the model's utilization of evidence documents in answering.

According to the above motivation and observations, we propose **CAFE**, a novel coarse-to-fine information-seeking method to enhance the multi-document question-answering capabilities of LLMs. Its core idea is to leverage the LLM's internal attention mechanisms to progressively identify and focus on question-relevant content in long contexts. Specifically, before information-seeking, we pre-locate the retrieval heads for the two stages, respectively, on the validation set. Then, in the first stage, we implement a coarse-grained filtering approach to filter out background documents. We identify relevant documents assigned with high attention scores in each pre-detected retrieval head and further rerank these documents according to the summed scores from all retrieval heads. In the second stage, we guide the model using a fine-grained steering approach. We utilize another set of retrieval heads to further select relevant documents from these reranked documents, and employ attention steering on the most relevant content to answer the final questions. In this way, we can guide the LLMs to gradually search for evidence documents in the long context input and utilize them to better answer the questions. Additionally, the whole method is training-free and applicable to a wide range of downstream tasks.

We conduct extensive experiments to evaluate the proposed CAFE method using various LLMs. The results demonstrate that our method consistently outperforms existing strong baselines across five benchmarks and three LLMs (*e.g.,* achieving an 11.4% relative performance improvement compared to the supervised fine-tuning method).

2 Related Work

Long-Context Utilization in Language Models. Although extended the context length of LLMs successfully (Dubey et al., 2024; Yang et al., 2024; Dong et al., 2024), they still face significant challenges (*e.g.,* long-term decay (Chen et al., 2024) and lost-in-the-middle (Liu et al., 2024)) in utilizing long contexts effectively for complex tasks. To enhance the long-context utilization capacities, attention-based methods leverage the property of attention heads and positional encodings, enlarging the attention scores of the key tokens over the long inputs (Wu et al., 2024; Gema et al., 2024). Different from previous methods, our work employs a training-free two-stage framework, which identifies relevant documents and guild the response more dependent on these documents.

Retrieval Head in Attention Mechanisms. Recent studies have revealed specialized attention heads in LLMs that exhibit retrieval capabilities for locating critical information within long contexts, namely, retrieval heads (Wu et al., 2024). In these heads, high attention values will be assigned to the tokens most relevant to the current token in the long inputs, achieving in-context retrieval of previous information. Recently, some work retains the full attention on the retrieval heads and employs KV Cache compression on other heads to accelerate the calculation (Fu et al., 2024a; Tang et al., 2024; Xiao et al., 2024). Different from them, our method utilizes retrieval heads as a retrieval system to identify evidence documents.

Retrieval-Augmented Generation. Retrieval-Augmented Generation (RAG) has been widely adopted to address various NLP tasks. For multi-document question-answering tasks, traditional RAG methods utilize external dense or sparse retrieval models to compute the similarity of documents with the question (Robertson and Zaragoza, 2009; Karpukhin et al., 2020). Then, relevant docu-

ments are retrieved as the input for models. Beyond leveraging external models to retrieve documents, several in-context retrieval methods have been proposed (Agrawal et al., 2024; Li et al., 2024a). These methods prompt the models to select the indices of relevant documents. Unlike existing RAG approaches, our work leverages the model's inherent retrieval capabilities to perform a coarse-to-fine location of evidence documents, effectively enhancing its retrieval and reasoning abilities.

3 Empirical Study

When dealing with multiple documents, humans often follow an identification–focusing–reasoning paradigm. Inspired by this, we conduct empirical studies to investigate an LLM-centric framework following this paradigm. Specifically, we analyze the effectiveness of various evidence selection strategies for identifying relevant documents, as well as the impact of information flow across different segments within attention modules.

3.1 Evidence Selection

As shown in Figure 1, the golden evidence is essential for multi-document question answering tasks. Thus, we first evaluate the effectiveness of evidence selection approaches in this scenario. We consider four primary methods: (1) *external retrieval models*: employing retrieval models to select documents; (2) *in-context retrieval(ICR)*: prompts LLMs to directly select documents most relevant to the question; (3) *attention-based retrieval*: employing the averaged attention scores of all heads over each document for selection; and (4) *retrieval head-based retrieval*: only employing the attention scores of retrieval heads for selection (Wu et al., 2024). Table 1 summarizes their recall across various datasets. Compared to external retrieval models and ICR, the attention-based approach significantly improves performance over ICR. Additionally, the retrieval head-based method further improves retrieval recall across all datasets evaluated. These results demonstrate that leveraging attention scores from specific retrieval heads provides a powerful and highly effective signal for relevant documents. More experimental results can be found in Appendix C.

3.2 Attention Intervention

To further explore the model's reasoning mechanism over multiple retrieved documents, we employ attention intervention techniques to adjust the information flow across different parts of the prompts. Specifically, we select test samples from HotpotQA-8K. Subsequently, we mask the attention between the two gold documents, as well as the attention from the question to the two gold documents and to the two irrelevant documents. We show the results in Table 2. First, masking the attention between the two gold documents has a negligible impact on performance compared to the unmasked condition. This suggests that during multi-hop question answering, the LLM does not engage in implicit reasoning[1] while encoding long inputs. Moreover, when applying attention mask between the question and irrelevant documents, we observe minimal performance impact. Conversely, masking the attention from the question to any gold document results in a significant performance drop. When all gold documents are masked simultaneously, the SubEM score even decreases to a level similar to that observed when no document is provided. This indicates that the information flow from the gold evidence to the question plays a crucial role in long-context QA performance. This motivates us to further explore ways to enhance the model's reasoning ability by selectively modifying this attention pathway.

Method	HQA-8k	HQA-32k	SQuAD	Musique
Retrieval Model	0.90	0.84	0.93	0.83
ICR	0.64	0.38	0.91	0.58
Attention	0.80	0.77	0.94	0.79
Retrieval Head	**0.97**	**0.93**	**0.97**	**0.9**

Table 1: Recall for different evidence selection strategies across four datasets using Llama-3.1-8B-Instruct.

Mask Mode	SubEM
No Mask	**60.5**
$\text{Evidence}_2 \rightarrow \text{Evidence}_1$	60.0
$\text{Question} \rightarrow \text{Evidence}_1$	48.0
$\text{Question} \rightarrow \text{Evidence}_2$	42.0
$\text{Question} \rightarrow \text{Evidence}_1, \text{Evidence}_2$	29.5
Question → One Irrelevant Document	60.0
Question → Two Irrelevant Documents	59.5

Table 2: SubEM scores on HotpotQA-8K with various masking strategies using Llama-3.1-8B-Instruct, where Evidence_1 and Evidence_2 refer to the first and second gold documents in the context.

Building on the prior experiments, we observe

[1]Implicit reasoning in our work refers to the model's ability to link information across documents before the final question is given. It encodes content from earlier documents into later ones through attention, enabling the question to be answered using only the later one.

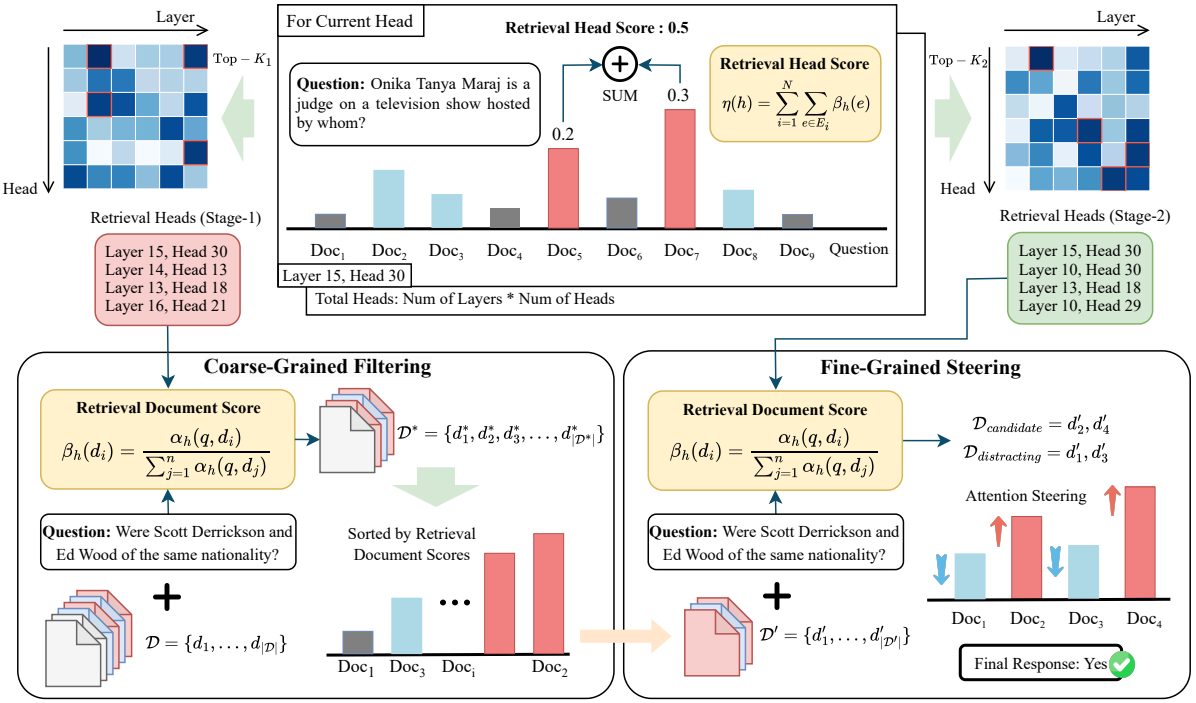

Figure 2: Overall framework of our proposed CAFE approach. The red, blue, and yellow bar charts represent the gold, distracting, and background documents, respectively.

that employing retrieval heads effectively extracts evidence documents from long inputs. Moreover, modifying the attention between the question and the evidence documents impacts the model's utilization of evidence documents for answering. These observations inform the design of our method.

4 Method

4.1 Overall Framework

In multi-document question-answering tasks, there are three categories of documents, *i.e.,* gold evidence documents that contain information supporting answering the questions, distracting documents that impede the model's ability to generate faithful answers, and background documents that contain irrelevant information. Among them, the latter two categories of documents can hardly be distinguished by simple retrieval. Inspired by human behaviors and observations in Section 3, we propose **CAFE**, a coarse-to-fine two-stage framework to enhance the long-context question-answering capacities by gradually eliminating the negative impacts of background and distracting documents. In our framework, we first apply coarse-grained filtering to identify relevant documents, and then use fine-grained attention steering to guide the LLMs to focus on documents with a higher likelihood of being gold evidence and perform reasoning on them. Specifically, we employ retrieval heads to locate relevant documents and identify these heads with different calibration datasets during the two stages. The overall illustration is shown in Figure 2.

4.2 Retrieval Head Detection

In Section 3, we observe that leveraging attention scores of retrieval heads can effectively identify evidence documents. Thus, we first locate retrieval heads that can be further employed to seek the relevant documents during the two stages.

Retrieval Document Scores. Based on the analysis in Section 3.2, we focus on the attention scores from the question to the contextual documents. Therefore, we first compute the **retrieval document score** $\beta_h(d_i)$ by analyzing attention weight scores between the question q and each document d_i:

$$\beta_h(d_i) = \frac{\alpha_h(q, d_i)}{\sum_{j=1}^{n} \alpha_h(q, d_j)}, \quad (1)$$

where $\alpha_h(q, d_i)$ represents the attention weight between the query q and document d_i for attention head h, and n is the total number of documents in the current sample.

Top-K Retrieval Heads Selection. To effectively identify retrieval heads, we select N samples from

the validation set and calculate a **retrieval head score** for each attention head h based on the evidence documents' retrieval document scores on these validation samples:

$$\eta(h) = \sum_{i=1}^{N} \sum_{e \in E_i} \beta_h(e), \tag{2}$$

where E_i is the set of evidence documents for the i-th sample. We then select the Top-K attention heads $\mathcal{H}_{\text{ret}}$ with the highest retrieval head scores from all attention heads $\mathcal{H}$ as the retrieval heads.

$$\mathcal{H}_{ret} = \text{Top-}K(\eta(h)), \;\; h \in \mathcal{H}. \tag{3}$$

Notably, during the coarse-grained filtering and fine-grained steering stages, we employ different validation sets and select different retrieval heads according to the properties of the two stages. The distinction between the two types of retrieval heads is detailed in Appendix D.

4.3 Coarse-Grained Filtering

In Figure 1, we observe that a large number of documents leads to significant performance degradation. Thus, we introduce a coarse-grained filtering stage to filter background documents and obtain a condensed input. Specifically, this stage consists of two steps: background document filtering and locality-based re-ranking.

Background Documents Filtering. To identify background documents, we first compute the retrieval document scores of each document on selected retrieval heads $\mathcal{H}_{ret}$. For each head h, we select Top-M_1 documents based on the retrieval document scores $\beta_h(d)$ from all documents $\mathcal{D}$ and consider them as relevant documents. Then, we perform a union operation on these documents to obtain the relevant document set $\mathcal{D}^*$ and drop the other documents.

$$\mathcal{D}^* = \bigcup_{h \in \mathcal{H}_{\text{ret}}} \text{Top-}M_1(\beta_h(d)), \;\; d \in \mathcal{D}. \tag{4}$$

Locality-Based Re-Ranking. When processing long context, LLMs usually demonstrate the property of locality and lost-in-the-middle (Liu et al., 2024; Su et al., 2024). This means when critical information for answering the question is located at the end of the long document, the model often performs better. Thus, after obtaining the filtered set of documents $\mathcal{D}^*$, we apply a **locality-based re-ranking** mechanism to rank these documents. For the filtered candidate document set $\mathcal{D}^*$, we compute the **document relevance score** $\gamma_h(d)$ for each document as the sum of retrieval document scores of all retrieval heads:

$$\gamma_h(d) = \sum_{h \in \mathcal{H}_{\text{ret}}} \beta_h(d), \;\; d \in \mathcal{D}^*. \tag{5}$$

Subsequently, documents with higher document relevance scores are positioned later in the sequence, ensuring that more attention will be focused on the documents that are more likely to contain critical evidence during the generation of responses. Finally, we obtain the filtered and reranked document sequence $\mathcal{D}'$ as the input of next stage:

$$\mathcal{D}' = \{d'_1, \ldots, d'_{|D^*|}\}, \forall i < j, \gamma_h(d_i) \leq \gamma_h(d_j). \tag{6}$$

4.4 Fine-Grained Steering

After the first stage of filtering irrelevant documents, most remaining documents are relevant to the question. However, there may still exist distracting documents. Thus, in the fine-grained steering stage, we further identify documents with a high likelihood of being golden evidence and steer the attention scores to guide the LLMs' attention towards these documents for better reasoning.

Iterative Distracting Document Identification. Similar to the coarse-grained filtering stage, to effectively identify and weaken the impact of these distracting documents, we perform document identification by computing retrieval document scores using another set of retrieval heads $\mathcal{H}'_{\text{ret}}$:

$$\mathcal{D}_{\text{cand}} = \bigcup_{h \in \mathcal{H}'_{\text{ret}}} \text{Top-}M_2(\beta_h(d)), \; d \in \mathcal{D}' \tag{7}$$

By identifying documents with high retrieval document scores, we ultimately derive a candidate set of evidence documents $\mathcal{D}_{\text{cand}}$. Each document in the candidate set is considered the golden evidence while other documents are considered as distracting documents during the following process of attention steering.

Inference-Time Attention Steering. After the initial filtering stage, the number of remaining documents is significantly reduced. In this stage, directly removing detected distractors may result in lower recall of evidence documents. Thus, instead of only keeping the candidate set, we adopt

post-hoc attention steering (Zhang et al., 2024a), an inference-only technique that reweights attention scores to guide the model's focus toward user-specified input spans. Specifically, given the candidate gold evidence set $\mathcal{D}_{\text{cand}}$, our method emphasizes specific tokens by adding a constant attention bias $\mathbf{B}^h$ to the attention scores on tokens within these documents across all attention heads.

$$\tilde{\mathbf{A}}^h = \text{Softmax}((\mathbf{Q}^{h\intercal}\mathbf{K}^h + \mathbf{B}^h)/\sqrt{d}), \quad (8)$$

$$B_{ij}^h = \begin{cases} \delta & \text{if } i \in q \text{ and } j \in \mathcal{C}_{\text{cand}} \\ 0 & \text{otherwise} \end{cases}, \quad (9)$$

where δ is a positive constant that controls the degree of attention adjustment. After applying Softmax$(\cdot)$, the attention scores of tokens in $\mathcal{D}_{\text{cand}}$ are enlarged while the attention scores of other tokens are reduced. This dynamic reweighting mechanism effectively enhances the model's attention toward tokens in $\mathcal{D}_{\text{cand}}$, ensuring the responses are more dependent on the critical evidence.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate the long-context performance of our approach and baseline methods using three question-answering datasets: SQuAD (Rajpurkar et al., 2016), HotpotQA (Yang et al., 2018), and MusiQue (Trivedi et al., 2022). These datasets are collected from the RULER (Hsieh et al., 2024) and LongBench (Bai et al., 2024a) benchmarks. Additionally, we experiment with three versions of HotpotQA that vary in context length to analyze how model performance changes with text length. To ensure consistency across all baselines and our approach, we randomly select 200 samples from each dataset to form the final test set. All experiments are conducted using the same test sets.

Baselines and Metrics. For evaluation, we use Substring Exact Match (SubEM) and F1 scores following existing work (Li et al., 2024b). SubEM measures whether the gold answer appears as a substring in the predictions, while the F1 score evaluates the token-level overlap between predictions and references. For compared baselines, we select five types of methods, including *Directly Answering*, *In-Context Retrieval*, *Oracle RAG*, *Vanilla RAG*, and *Supervised Fine-tuning*. We present the detailed description in Appendix B.

Implementation Details. We conduct our experiments on three open-source models: Llama-3.1-8B-Instruct, Mistral-3-7B-Instruct, and Phi-3.5-Mini-Instruct. For coarse-grained filtering for background documents, we set the Top-M_1 to 4 and Top-K_1 to 4. For fine-grained steering for distracting text, we set the Top-M_2 to 2 and Top-K_2 to 2 and we set $\delta = \log 10$. As for the SFT configuration, training is conducted with a batch size of 64 and a learning rate of 1×10^{-5} for 1 epoch.

5.2 Main Results

Table 3 shows the results of our methods and other baselines across three representative long context question-answering datasets.

Firstly, our method achieves significantly better multi-document question-answering performances than other baselines. Across all three datasets, our method consistently outperforms training-free approaches and even surpasses the SFT method in most settings. On single-hop SQuAD, our method can achieve performances nearly the performance ceiling introduced by Oracle RAG. On more complex multi-hop question-answering tasks, our method can still achieve a significant performance improvement (*e.g.,* approximately 19.9% of SubEM scores on the HotpotQA dataset compared to the naive directly answering method).

Secondly, the two stages of our method work together to prompt performance improvements. Compared with in-context retrieval and vanilla RAG which retrieve relevant documents via prompting techniques or external models, only employing the coarse-grained filtering stage can greatly boost the performance, indicating that leveraging the inner retrieval heads can more effectively identify relevant documents. Additionally, introducing fine-grained attention steering can further boost long-context question-answering capacities, which demonstrates the necessity of introducing a fine-grained elimination of the negative impacts of distracting documents on multi-document question answering.

Finally, our method exhibits less performance drop with longer input lengths. On the HotpotQA dataset, we assess the performances across different input lengths. Our method can preserve performance to a greater extent when dealing with longer texts (*e.g.,* decreases 1.4% and 2.1% SubEM scores for Llama-3.1-8B on 16K and 32K). Instead, the performances drop sharply with the length increasing with other methods, especially in-context retrieval (*e.g.,* decreases 12.7% and 28.0% SubEM

LCLM	Baseline	SQuAD		MuSiQue		HotpotQA		HotpotQA-16K		HotpotQA-32K	
		SubEM	F1	SubEM	F1	SubEM	F1	SubEM	F1	SubEM	F1
Llama-3.1-8B	Oracle RAG	92.5	86.4	39.0	39.3	76.5	76.8	76.5	76.8	76.5	76.8
	Directly Answering	71.0	66.6	30.5	33.2	60.5	62.5	53.0	60.1	53.5	58.1
	In-Context Retrieval	73.5	65.1	28.0	29.2	59.0	58.6	51.5	51.8	42.5	42.2
	Vanilla RAG	84.5	76.6	28.0	29.7	64.0	64.8	63.0	63.6	61.5	62.4
	SFT	69.0	70.1	33.5	**38.9**	63.0	69.8	62.5	68.0	61.5	67.4
	CAFE (w/o FGS)	89.5	80.7	36.0	35.5	68.5	69.0	66.0	68.3	66.0	65.2
	CAFE (ours)	**89.5**	**82.6**	**36.5**	36.5	**70.0**	**70.4**	**69.0**	**69.0**	**68.5**	**68.1**
Mistral-3-7B	Oracle RAG	84.0	80.1	40.5	38.9	67.0	71.3	67.0	71.3	67.0	71.3
	Directly Answering	59.0	55.9	27.5	26.8	50.0	53.7	45.0	47.5	39.0	46.6
	In-Context Retrieval	59.5	58.7	24.0	24.2	49.0	47.6	37.5	38.2	29.5	30.3
	Vanilla RAG	69.5	69.2	27.5	26.2	53.5	55.9	53.5	55.4	51.0	54.7
	SFT	60.0	60.1	30.5	**33.1**	57.5	61.9	52.5	56.7	47.5	53.6
	CAFE (w/o FGS)	78.0	73.6	30.0	27.9	60.0	64.0	60.5	60.0	53.0	56.5
	CAFE (ours)	**78.5**	**75.2**	**31.0**	29.9	**61.5**	**65.2**	**60.5**	**61.7**	**58.0**	**61.7**
Phi-3.5-Mini	Oracle RAG	85.0	80.0	35.0	38.1	73.0	75.8	73.0	75.8	73.0	75.8
	Directly Answering	63.5	58.8	24.5	27.5	55.0	55.5	51.5	52.5	48.0	48.3
	In-Context Retrieval	65.5	66.4	22.5	23.7	49.5	49.5	38.0	39.5	31.0	34.4
	Vanilla RAG	76.0	72.5	25.5	26.1	58.5	60.2	56.0	58.8	55.0	58.7
	SFT	64.5	65.1	**34.5**	**40.9**	60.5	**71.8**	61.0	**71.8**	58.0	**67.3**
	CAFE (w/o FGS)	82.0	74.9	28.5	28.8	65.0	67.8	64.5	62.6	60.0	58.9
	CAFE (ours)	**84.5**	**75.8**	30.0	31.9	**66.5**	68.0	**66.5**	64.8	**61.5**	60.1

Table 3: Evaluation results on three long-document question answering tasks. They are representative of single-hop and multi-hop question-answering tasks. "CAFE (w/o FGS)" means that we only perform the first stage without the fine-grained steering for the distracting text stage. The **bold** and underline fonts denote the best and second-best results in each dataset. Notably, all models in the table are the instructed versions.

scores for Llama-3.1-8B on 16K and 32K). This indicates that our method can effectively identify the critical documents in the long input, scarcely affected by the increased number of documents.

5.3 Further Analysis

Ablation Study. To assess the effectiveness of our framework, we conduct ablation experiments focusing on key steps within the pipeline. (1) *w/o Coarse-Grained Filtering (CGF)* eliminates the initial coarse-grained filtering of background documents; (2) *w/o Fine-Grained Steering (FGS)* omits the fine-grained steering of distracting text, relying solely on documents $\mathcal{D}'$ for inference; (3) *w/o Locality-Based Re-Ranking* bypasses locality-based re-ranking in the first stage, resulting in the use of filtered documents in a random order.

Method	Llama	Mistral	Phi
CAFE	70.0	61.5	66.5
w/o CGF	62.5	52.0	55.0
w/o FGS	68.5	60.0	65.0
w/o Re-Ranking	68.0	59.5	65.5

Table 4: Ablation study on HotpotQA.

The results are presented in Table 4. All variants show inferior performance compared to the original method, underscoring the effectiveness of each component in our framework. Notably, the absence of Coarse-Grained Filtering (*w/o CGF*) results in a substantial performance decline, highlighting the critical role of first-stage filtering in excluding irrelevant background documents and preventing the dilution of the model's attention. Similarly, the removal of Fine-Grained Steering (*w/o FGS*) leads to decreased performance, indicating that the second stage's attention steering effectively mitigates the impact of distracting documents. Furthermore, the exclusion of Re-Ranking (*w/o Re-Ranking*) results in significant performance degradation, demonstrating the effectiveness of putting the essential information at the end of the input to facilitate the retrieval and reasoning of models.

Lost-in-the-Middle Performance. We investigate the Lost-in-the-Middle phenomenon and the effectiveness of our method in mitigating it. Experiments are conducted on the SQuAD dataset using LLaMA and Mistral, evaluating how the position of the answer within a set of 50 documents af-

Model	Method	1	10	20	30	40	50	Rand
LLaMA-3.1-8B-Instruct	DA	77.5/70.9	74.5/67.4	73.0/67.6	70.5/64.6	69.5/64.4	73.0/67.6	71.0/66.6
	Ours	**90.5/82.1**	**91.0/80.3**	**89.5/80.9**	**89.0/80.5**	**88.5/79.9**	**89.5/79.2**	**89.5/82.6**
Mistral-3-7B-Instruct	DA	70.0/58.7	59.0/50.6	56.5/48.3	59.5/51.6	58.0/52.9	62.0/59.5	59.0/55.9
	Ours	**79.5/76.0**	**80.0/74.8**	**79.0/71.9**	**78.5/71.1**	**78.5/71.2**	**78.0/72.6**	**78.5/75.2**

Table 5: Position-wise SubEM/F1 scores on two models. The column headers (1, 10, 20, *etc*) indicate the document index where the gold document is inserted. DA denotes Directly Answering.

fects model performance. As shown in Table 5, the Lost-in-the-Middle phenomenon significantly degrades the baseline method's performance, particularly when answers are in middle positions (e.g., Mistral's SubEM score drops from 70% to 58%). Our method effectively mitigates the issue, achieving stable and significantly improved performance across all answer positions, consistently outperforming the baseline. This approach demonstrates strong robustness and generalizability, requiring no position-specific adjustments.

Granularity of Attention Steering. In the fine-grained steering stage, we also evaluate the impact of the granularity of attention steering. Instead of document-level, we identify relevant contexts at the sentence level and steer the attention scores on these sentences. As shown in Table 6, the recall at the sentence level is lower compared to the document level. Additionally, the final performances degrade significantly, even inferior to those before attention steering. This indicates the importance of covering the golden evidence information as much as possible during the attention steering stages.

Granularity	Recall	SubEM	F1
w/o Steering	-	68.5	69.0
Sentence-Level	0.89	65.5	67.8
Document-Level	0.93	70.0	70.4

Table 6: Results with different steering granularities.

Impact of Hyperparameters. The choice of hyperparameters M (documents per head) and K (number of heads) during retrieval head selection has a strong influence on both recall and overall performance. As shown in Figure 3, increasing M_1 or K_1 boosts recall by adding more candidates, but can also introduce noise that limits final accuracy. We therefore fix $M_1 = 4$ and $K_1 = 4$ for consistency. A similar trade-off holds in the second stage, though performance remains stable after attention steering. The remaining hyperparameter details are provided in Appendix E.

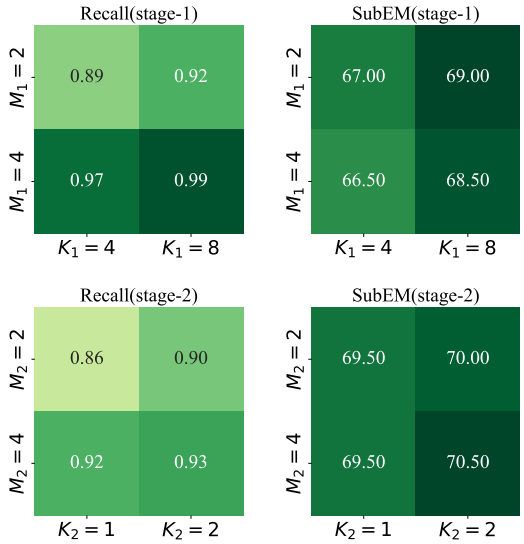

Figure 3: The impact of hyperparameters M (documents per retrieval head) and K (retrieval heads) on Llama-3.1-8B-Instruct. The top row shows recall and performance for coarse-grained filtering, while the bottom row illustrates changes for fine-grained steering.

6 Conclusion

In this paper, we explored the challenges faced by LLMs in handling long-context inputs, particularly in multi-document question answering tasks. Our findings revealed that the inclusion of irrelevant documents significantly hampers the retrieval and reasoning capabilities of LLMs, motivating the need for more effective long-context processing strategies. To address this, we introduced CAFE, a two-stage coarse-to-fine information-seeking method that leverages retrieval head-based filtering, document reranking, and fine-grained attention steering to guide LLMs in processing long-context inputs. Extensive experiments across multiple benchmarks and LLMs validate its effectiveness, demonstrating its superiority over strong baselines, including supervised fine-tuning techniques. Beyond its performance benefits, CAFE's training-free nature and broad applicability make it a practical solution for a wide range of downstream tasks.

Limitations

In this paper, we present a coarse-to-fine two-stage framework to enhance the retrieval and reasoning capacities of LLMs. Beyond multi-document question answering tasks, we believe our framework can be employed in broader tasks, *e.g.,* long-document reasoning, which have not been explored owing to the computational costs. Additionally, our method mainly focuses on how to better identify evidence documents to enhance performance. However, though given the golden evidence, the LLMs can still hardly answer each question correctly. Approaches to improving the context-aware reasoning capacities can be employed to further improve the upper limit of our method.

References

Devanshu Agrawal, Shang Gao, and Martin Gajek. 2024. Can't remember details in long documents? you need some r&r. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 12692–12704. Association for Computational Linguistics.

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2024. Make your LLM fully utilize the context. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024a. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 3119–3137. Association for Computational Linguistics.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024b. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *CoRR*, abs/2412.15204.

Yuhan Chen, Ang Lv, Jian Luan, Bin Wang, and Wei Liu. 2024. Hope: A novel positional encoding without long-term decay for enhanced context awareness and extrapolation. *CoRR*, abs/2410.21216.

Zican Dong, Junyi Li, Xin Men, Xin Zhao, Bingning Wang, Zhen Tian, Weipeng Chen, and Ji-Rong Wen. 2024. Exploring context window of large language models via decomposed positional vectors. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Tianyu Fu, Haofeng Huang, Xuefei Ning, Genghan Zhang, Boju Chen, Tianqi Wu, Hongyi Wang, Zixiao Huang, Shiyao Li, Shengen Yan, Guohao Dai, Huazhong Yang, and Yu Wang. 2024a. Moa: Mixture of sparse attention for automatic large language model compression. *CoRR*, abs/2406.14909.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. 2024b. Data engineering for scaling language models to 128k context. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Aryo Pradipta Gema, Chen Jin, Ahmed Abdulaal, Tom Diethe, Philip Teare, Beatrice Alex, Pasquale Minervini, and Amrutha Saseendran. 2024. Decore: Decoding by contrasting retrieval heads to mitigate hallucinations. *CoRR*, abs/2410.18860.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. RULER: what's the real context size of your long-context language models? *CoRR*, abs/2404.06654.

Jinhao Jiang, Jiayi Chen, Junyi Li, Ruiyang Ren, Shijie Wang, Wayne Xin Zhao, Yang Song, and Tao Zhang. 2024. Rag-star: Enhancing deliberative reasoning with retrieval augmented verification and refinement. *CoRR*, abs/2412.12881.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. 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 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

Jinhyuk Lee, Anthony Chen, Zhuyun Dai, Dheeru Dua, Devendra Singh Sachan, Michael Boratko, Yi Luan, Sébastien M. R. Arnold, Vincent Perot, Siddharth Dalmia, Hexiang Hu, Xudong Lin, Panupong Pasupat, Aida Amini, Jeremy R. Cole, Sebastian Riedel, Iftekhar Naim, Ming-Wei Chang, and Kelvin Guu. 2024. Can long-context language models subsume retrieval, rag, sql, and more? *CoRR*, abs/2406.13121.

Huayang Li, Pat Verga, Priyanka Sen, Bowen Yang, Vijay Viswanathan, Patrick Lewis, Taro Watanabe, and Yixuan Su. 2024a. Alr^2: A retrieve-then-reason framework for long-context question answering. *CoRR*, abs/2410.03227.

Siheng Li, Cheng Yang, Zesen Cheng, Lemao Liu, Mo Yu, Yujiu Yang, and Wai Lam. 2024b. Large language models can self-improve in long-context reasoning. *CoRR*, abs/2411.08147.

Yanyang Li, Shuo Liang, Michael R. Lyu, and Liwei Wang. 2024c. Making long-context language models better multi-hop reasoners. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 2462–2475. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguistics*, 12.

Tsendsuren Munkhdalai, Manaal Faruqui, and Siddharth Gopal. 2024. Leave no context behind: Efficient infinite context transformers with infini-attention. *CoRR*, abs/2404.07143.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, pages 333–389.

Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Hanlin Tang, Yang Lin, Jing Lin, Qingsen Han, Shikuan Hong, Yiwu Yao, and Gongyi Wang. 2024. Razorattention: Efficient KV cache compression through retrieval heads. *CoRR*, abs/2407.15891.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. 9835 musique: Multi-hop questions via single-hop question composition. *Trans. Assoc. Comput. Linguistics*, 10.

Minzheng Wang, Longze Chen, Fu Cheng, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, Yunshui Li, Min Yang, Fei Huang, and Yongbin Li. 2024. Leave no document behind: Benchmarking long-context llms with extended multi-doc QA. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5627–5646. Association for Computational Linguistics.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2024. Retrieval head mechanistically explains long-context factuality. *CoRR*, abs/2404.15574.

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. 2024. Duoattention: Efficient long-context LLM inference with retrieval and streaming heads. *CoRR*, abs/2410.10819.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale,

Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. 2024. Effective long-context scaling of foundation models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 4643–4663. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics.

Howard Yen, Tianyu Gao, and Danqi Chen. 2024. Long-context language modeling with parallel context encoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 2588–2610. Association for Computational Linguistics.

Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. 2024a. Tell your model where to attend: Post-hoc attention steering for llms. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Qingru Zhang, Xiaodong Yu, Chandan Singh, Xiaodong Liu, Liyuan Liu, Jianfeng Gao, Tuo Zhao, Dan Roth, and Hao Cheng. 2024b. Model tells itself where to attend: Faithfulness meets automatic attention steering. *CoRR*, abs/2409.10790.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2024c. ∞bench: Extending long context evaluation beyond 100k tokens. *CoRR*, abs/2402.13718.

Andrew Zhu, Alyssa Hwang, Liam Dugan, and Chris Callison-Burch. 2024. Fanoutqa: A multi-hop, multi-document question answering benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024 - Short Papers, Bangkok, Thailand, August 11-16, 2024*, pages 18–37. Association for Computational Linguistics.

A Performance Gap

We study how distracting documents affects model performance in long-context settings. Starting with only gold evidence, we gradually insert irrelevant documents and observe a performance drop as shown in Figure 1. This suggests that longer inputs with more irrelevant content weaken the model's retrieval and reasoning. This motivates our design of a retrieval strategy to filter out such noise.

B Baselines

We compare CAFE with the following baselines:

• **Directly Answering.** Asking LLMs to directly answer the question by using the context.

• **In-Context Retrieval.** LLMs are initially prompted to generate the key documents that support answering the question. Then, models are prompted to answer the question with the key documents appended to the context.

• **Oracle RAG.** Asking LLMs to answer the question only based on the ground-truth documents to estimate an upper limit performance.

• **Supervised Fine-tuning.** The LLM is trained on training sets of these datasets. We randomly sample 2000, 5000, and 5000 training instances for SQuAD, HotpotQA, and MusiQue, respectively.

C Evidence Selection Results

We conduct additional validations on more models, and the results are shown in Table 8.

D Differences Between Coarse and Fine Retrieval Heads

To better understand the behavior of retrieval heads used in the two stages, we visualize their accumulated attention scores over documents in Figure 4 and Figure 5. We observe that coarse-stage retrieval heads focus on a few documents to filter background information, while fine-stage heads attend more broadly to distinguish gold evidence from distractors.

E Impact of Hyperparameters

We perform an ablation study on the fine-grained retrieval parameter δ. As shown in Table 7, values around $\log 10$ yield stable performance, and we use $\log 13$ as the default setting in our main experiments.

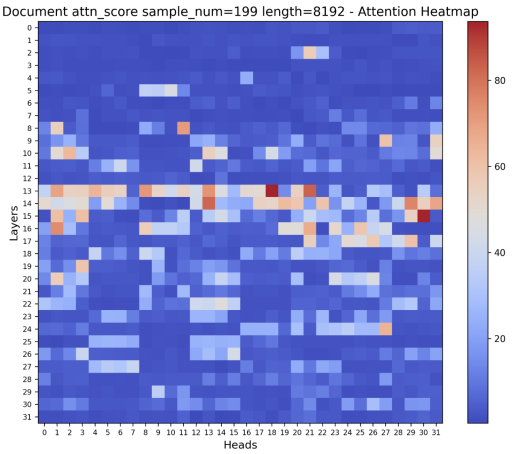

Figure 4: Attention distribution of retrieval heads used in the coarse-grained filtering stage.

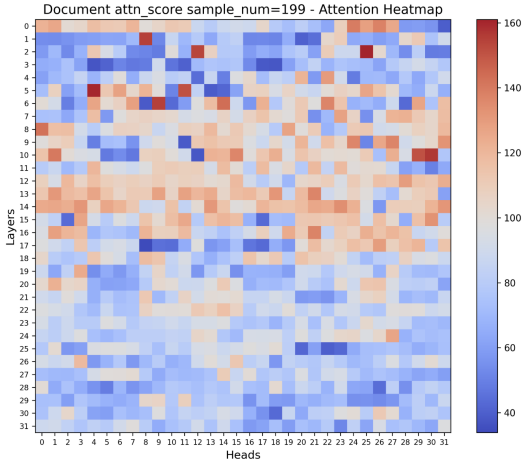

Figure 5: Attention distribution of retrieval heads used in the fine-grained steering stage.

F Inference Latency Evaluation

Our method requires multiple inferences (at least two prefilling operations), which indeed increases inference latency. In the first round, the prefill length is the same as that of direct answer. As a result, our method is slightly slower than the native flash-attention used in direct answer. We report the difference in prefill efficiency between our method and the Directly Answering baseline in Table 9.

δ	$\log 5$	$\log 10$	$\log 13$	$\log 20$
EM score	69.0	**70.5**	70.0	69.5

Table 7: Ablation study on the parameter δ.

Model	Method	HotpotQA-8k	HotpotQA-16k	HotpotQA-32k	SQuAD	Musique
Llama-3.1-70B-Instruct	ICR	0.82	0.72	0.54	0.92	0.67
	Attention-Based	0.85	0.81	0.77	0.95	0.81
Mistral-3-7B-Instruct	ICR	0.65	0.49	0.33	0.78	0.46
	Attention-Based	0.75	0.71	0.64	0.82	0.60
Phi-3.5-Mini-Instruct	ICR	0.39	0.27	0.21	0.72	0.33
	Attention-Based	0.54	0.52	0.43	0.73	0.36

Table 8: SubEM scores comparing In-Context Retrieval (ICR) and Attention-Based Retrieval methods across different context lengths and datasets. The second column indicates the retrieval method; performance declines for ICR as context length grows, while the attention-based approach remains more robust.

Method	HotpotQA-8k	HotpotQA-16k	HotpotQA-32k	SQuAD	Musique
Directly Answering	637.02	1482.25	2867.17	592.65	1493.73
Ours (prefill)	840.95	1799.92	3880.35	719.55	1791.15

Table 9: TTFT (ms/token) comparison across datasets. "Ours (prefill)" refers to the inference time including the prefill enhancement.