Efficient OpAmp Adaptation for Zoom Attention to Golden Contexts

Haoyuan Wu[♠*], **Rui Ming**[♠*], **Haisheng Zheng**[♡], **Zhuolun He**[♠,♣], **Bei Yu**[♠]

[♠]The Chinese University of Hong Kong, Hong Kong SAR
[♡]Shanghai Artificial Intelligent Laboratory, China
[♣]ChatEDA Tech, China
{hywu24,byu}@cse.cuhk.edu.hk

Abstract

Large language models (LLMs) have shown significant promise in question-answering (QA) tasks, particularly in retrieval-augmented generation (RAG) scenarios and long-context applications. However, their performance is hindered by noisy reference documents, which often distract from essential information. Despite fine-tuning efforts, Transformer-based architectures struggle to prioritize relevant content. This is evidenced by their tendency to allocate disproportionate attention to irrelevant or later-positioned documents. Recent work proposes the differential attention mechanism to address this issue, but this mechanism is limited by an unsuitable common-mode rejection ratio (CMRR) and high computational costs. Inspired by the operational amplifier (OpAmp), we propose the OpAmp adaptation to address these challenges, which is implemented with adapters efficiently. By integrating the adapter into pre-trained Transformer blocks, our approach enhances focus on the golden context without costly training from scratch. Empirical evaluations on noisy-context benchmarks reveal that our Qwen2.5-OpAmp-72B model, trained with our OpAmp adaptation, surpasses the performance of state-of-the-art LLMs, including DeepSeek-V3 and GPT-4o. Our code is available at https://github.com/wuhy68/OpampAdapter.

1 Introduction

Recent advancements in large language models (LLMs) (OpenAI, 2023; Dubey et al., 2024; Yang et al., 2024; Liu et al., 2024a) have demonstrated remarkable capabilities in understanding, generating, and reasoning across diverse domains, significantly advancing their application in various fields. Among these applications, question answering (QA) based on provided contexts has emerged as one of the most prominent use cases for LLMs.

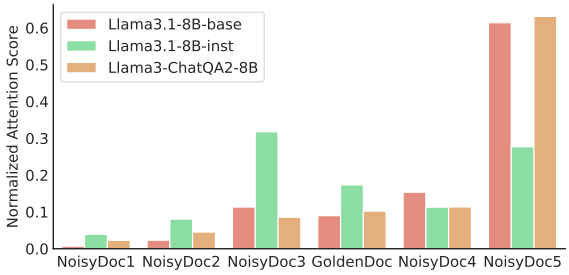

Figure 1: Normalized attention score. Transformers often miss the golden document in a noisy context.

As LLMs' capabilities continue to evolve and user expectations grow, users increasingly supply multiple documents retrieved in Retrieval-Augmented Generation (RAG) scenarios or long-context reference documents to guide LLMs in generating contextually relevant responses. However, in practice, such retrieved documents or long-context references often contain substantial noise, including information irrelevant to the user's query. Recent studies (Ye et al., 2025; Liu et al., 2024b) highlight a critical challenge that LLMs frequently struggle to accurately identify and extract key information from these noisy contexts, limiting their effectiveness in real-world applications.

As illustrated in Figure 1, we visualize the normalized attention scores assigned to retrieved documents in the RAG scenario, which includes various noisy documents and a single golden document. The task involves identifying the correct answer within noisy contexts. Our analysis evaluates several LLMs, including Llama3.1-8B-base (Meta, 2024), Llama3.1-8B-inst (Meta, 2024), and Llama3-ChatQA2-8B (Xu et al., 2024), the latter of which has been fine-tuned specifically for long-context and RAG applications. The visualization demonstrates that the Transformer architecture tends to allocate only a small proportion of attention scores to the golden document, while disproportionately focusing on irrelevant or

[*]These authors contributed equally to this work.

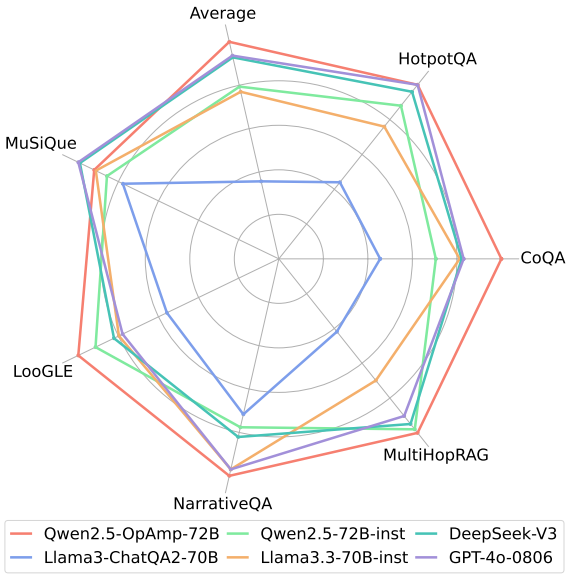

Figure 2: Qwen2.5-OpAmp-72B achieves the best average performance in various noisy-context benchmarks compared to current SOTA LLMs.

later-positioned documents. Notably, ChatQA2, despite its fine-tuning for long-context and RAG tasks, tends to over-attend to documents positioned later in the sequence rather than the golden document. Similarly, the aligned LLM struggles to focus on relevant information in noisy environments. These findings highlight a persistent challenge for Transformer-based architectures, including effectively identifying and prioritizing relevant documents in the presence of noise. The issue (Ye et al., 2025) arises from the non-negligible allocation of attention scores to irrelevant content, which ultimately obscures the correct answer and undermines model performance.

Ye et al. (2025) propose a differential attention mechanism designed to mitigate attention noise through differential denoising, inspired by the principles of differential amplifiers in electrical engineering. However, differential amplifiers are effective in scenarios requiring a high common-mode rejection ratio (CMRR) considering that they only focus on differential gain. This is unsuitable for attention denoising in the Transformer block. Training a differential transformer from scratch entails great computation costs and introduces significant risks, further limiting its practical applicability.

Inspired by the operational amplifiers (OpAmp), we introduce OpAmp adaptation with adapters, an efficient approach for refining the attention mechanism to enhance focus on the most relevant context leveraging parameter-efficient fine-tuning (PEFT) techniques. The OpAmp adaptation enables simultaneous control of differential gain and common-mode gain through the management of the CMRR. Building on the OpAmp design, our approach facilitates the training of OpAmp models using pre-trained Transformer architectures, eliminating the need for training from scratch. This strategy significantly reduces computational costs compared to previous methods. As demonstrated in Figure 2, our Qwen2.5-OpAmp-72B model, trained with the OpAmp adaptation, achieves superior average performance across various noisy-context benchmarks compared to current state-of-the-art (SOTA) LLMs. Our contributions are as follows:

- We introduce the OpAmp adaptation for zoom attention to the most relevant context in noisy contexts;
- Implement OpAmp adaptation with adapters, which are fine-tuned with our noisy context dataset, achieving significant improvements;
- Develop OpAmp models with our OpAmp adaptation method, surpassing current SOTA LLMs in various noisy-context benchmarks.

2 Methods

2.1 Preliminaries

Adapters. Houlsby et al. (2019) introduced the concept of integrating adapters into pre-trained transformer-based models for PEFT. This approach only fine-tunes the parameters introduced by the adapters while maintaining the pre-trained weights with large parameters unchanged. An adapter module comprises two trainable matrices, $\boldsymbol{W}_1 \in \mathbb{R}^{d_1 \times d_2}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{d_2 \times d_1}$, along with a non-linear activation function $\phi(\cdot)$. Here, d_1 represents the feature dimension of the pre-trained weights, while d_2 denotes the hidden dimension of the inserted adapter, typically satisfying $d_2 \ll d_1$. Given an input feature $\boldsymbol{H} \in \mathbb{R}^{N \times d_1}$, the output of the adapter module is expressed as:

$$\boldsymbol{H}' = \phi(\boldsymbol{H}\boldsymbol{W}_1)\boldsymbol{W}_2 + \boldsymbol{H}. \quad (1)$$

Attention. The self-attention mechanism (Vaswani et al., 2017) serves as the foundational building block for LLMs (OpenAI, 2023; Dubey et al., 2024; Yang et al., 2024; Liu et al., 2024a). Given a query feature $\boldsymbol{Q} \in \mathbb{R}^{N \times d}$, a key feature $\boldsymbol{K} \in \mathbb{R}^{N \times d}$, and a value feature $\boldsymbol{V} \in \mathbb{R}^{N \times d}$, the attention mecha-

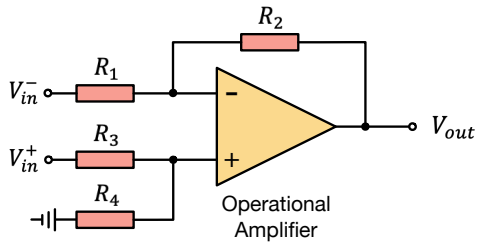

Figure 3: The operational amplifier with two input voltages V_{in}^+ and V_{in}^-. The CMRR $\mathcal{K}$ is controlled by resistances R_1, R_2, R_3, R_4.

nism is computed as follows:

$$\text{Attn}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \boldsymbol{M}\boldsymbol{V},$$
$$\boldsymbol{M} = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d}}\right), \quad (2)$$

where N represents the number of tokens and d denotes the dimensionality of the query, key, and value features.

Differential Amplifier. The differential amplifier (Sansen, 2007) is an electronic device designed to amplify the voltage difference between its two input signals while rejecting any voltage common to both inputs. In an analog circuit with input voltages V_{in}^+ and V_{in}^-, the ideal output voltage V_{out} is proportional to the difference between the two inputs, as expressed by:

$$V_{\text{out}} = A_d(V_{\text{in}}^+ - V_{\text{in}}^-), \quad (3)$$

where A_d represents the differential gain.

Operational Amplifier. In practical applications, the desired output often deviates from the predictions of Equation (3). For instance, when V_{in}^+ and V_{in}^- are equal, the output voltage does not necessarily become zero. However, according to Equation (3), the output voltage should theoretically be zero in such cases. To address this discrepancy, as shown in Figure 3, the OpAmp (Sansen, 2007) provides a more accurate and stable output expression, including an additional term accounting for common-mode effects:

$$\begin{aligned} V_{\text{out}} &= V_{\text{in}}^+ \cdot \left(\frac{R_4}{R_3 + R_4} \cdot \frac{R_1 + R_2}{R_1}\right) - V_{\text{in}}^- \cdot \frac{R_2}{R_1} \\ &= A_d(V_{\text{in}}^+ - V_{\text{in}}^-) + \frac{A_c}{2}(V_{\text{in}}^+ + V_{\text{in}}^-), \end{aligned} \quad (4)$$

where A_c is the common-mode gain of the amplifier. The common-mode rejection ratio (CMRR) is defined as the ratio of the differential gain to the common-mode gain:

$$\mathcal{K} = \frac{A_d}{A_c}. \quad (5)$$

Obviously, $A_c \to 0$ and $\mathcal{K} \to \infty$ for an ideal differential amplifier.

2.2 OpAmp Adaptation

Inspired by the operational amplifier, we propose the OpAmp adaptation, which modifies the original attention mechanism into the OpAmp attention mechanism. Specifically, the operational amplifier is employed to denoise the input signals and produce a refined output in the analog circuit domain. Building on this concept, we design the OpAmp attention mechanism to denoise the attention matrices $\boldsymbol{M}$. As shown in Figure 4, the original attention mechanism described in Equation (2) is adapted using Equation (4):

$$\bar{\boldsymbol{M}} = A_d(\boldsymbol{M}^+ - \boldsymbol{M}^-) + \frac{A_c}{2}(\boldsymbol{M}^+ + \boldsymbol{M}^-), \quad (6)$$

where $\bar{\boldsymbol{M}}$ is the denoised attention matrix via OpAmp adaptation, $\boldsymbol{M}^+$ and $\boldsymbol{M}^-$ are formulated through adapters, the detailed implementation of which will be elaborated in Section 2.3. As illustrated in Equation (6), we can adopt different $\mathcal{K}$ to adapt different scenarios using Equation (5).

Notably, the attention noise for LLMs after alignment is relatively small in noisy-context scenarios as shown in Figure 1. This suggests that attention denoising requires only a modest CMRR $\mathcal{K}$ instead of high CMRR values. The experiment results presented in Section 3.4 further support our claim that excessively high CMRR values can lead to performance degradation.

2.3 Architecture Design

Given an input feature $\boldsymbol{X} \in \mathbb{R}^{N\times d}$, the query feature $\boldsymbol{Q} \in \mathbb{R}^{N\times d}$ and the key feature $\boldsymbol{K} \in \mathbb{R}^{N\times d}$ are computed as follows:

$$\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}^q, \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}^k, \quad (7)$$

where $\boldsymbol{W}^q, \boldsymbol{W}^k \in \mathbb{R}^{d\times d}$ represent pre-trained weights used for linear projection. As outlined in Equation (6), the computation of $\boldsymbol{M}^+$ and $\boldsymbol{M}^-$ is required to implement the OpAmp attention mechanism. A straightforward approach involves duplicating $\boldsymbol{W}^Q$ and $\boldsymbol{W}^K$ to compute two sets of query and key features, denoted as $\boldsymbol{Q}_1, \boldsymbol{K}_1$ and $\boldsymbol{Q}_2, \boldsymbol{K}_2$

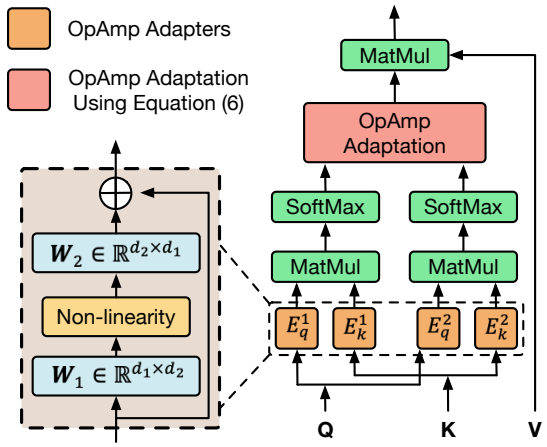

Figure 4: Overview of the OpAmp adaptation using Equation (6) with adapters.

Subsequently, $\boldsymbol{M}^+$ and $\boldsymbol{M}^-$ can be calculated independently using Equation (2) as follows:

$$\boldsymbol{M}^+ = \text{Softmax}\left(\frac{\boldsymbol{Q}_1\boldsymbol{K}_1^\top}{\sqrt{d}}\right), \tag{8}$$

$$\boldsymbol{M}^- = \text{Softmax}\left(\frac{\boldsymbol{Q}_2\boldsymbol{K}_2^\top}{\sqrt{d}}\right), \tag{9}$$

However, this method incurs substantial computational overhead, particularly given the large parameter scale of LLMs.

Consequently, we introduce an effective and efficient implementation of OpAmp adaptation to address this inefficiency. Specifically, we employ adapters to avoid redundant weight computations as shown in Figure 4. For a given input $\boldsymbol{X}$, the query and key features $\boldsymbol{Q}_1, \boldsymbol{K}_1$ and $\boldsymbol{Q}_2, \boldsymbol{K}_2$ can be computed as follows:

$$\boldsymbol{Q}_1 = E_q^1(\boldsymbol{X}\boldsymbol{W}^q), \boldsymbol{Q}_2 = E_q^2(\boldsymbol{X}\boldsymbol{W}^q), \tag{10}$$

$$\boldsymbol{K}_1 = E_k^1(\boldsymbol{X}\boldsymbol{W}^k), \boldsymbol{K}_2 = E_k^2(\boldsymbol{X}\boldsymbol{W}^k), \tag{11}$$

where $E_j^i(\boldsymbol{x})$ represents the adapters for OpAmp adaptation, defined according to Equation (1) as:

$$E_j^i(\boldsymbol{x}) = \phi(\boldsymbol{x}\boldsymbol{W}_1)\boldsymbol{W}_2 + \boldsymbol{x}, \tag{12}$$

with $i \in \{1, 2\}$ and $j \in \{q, k\}$. This architecture ensures effective OpAmp adaptation while minimizing computational overhead. Finally, the output of OpAmp attention can be computed as:

$$\text{OpAmpAttn}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \bar{\boldsymbol{M}}\boldsymbol{V}. \tag{13}$$

Zero Initialization. At the onset of training, we employ zero initialization to promote identity mapping. Specifically, $\boldsymbol{W}_2$ is initialized to zero to guarantee that $E_j^i(\boldsymbol{x}) = \boldsymbol{x}$. Furthermore, to prevent any disruption to the original $\boldsymbol{M}$ during the initial phase of training, we set $A_c = 1$ and regulate $\mathcal{K} = \frac{A_d}{A_c}$ by adjusting the values of A_d. As a result, at the initial stage, Equation (6) reduces to:

$$\begin{aligned}\bar{\boldsymbol{M}} &= A_d \cdot (\boldsymbol{M} - \boldsymbol{M}) + \frac{A_c}{2} \cdot (\boldsymbol{M} + \boldsymbol{M}), \\ &= A_d \cdot 0 + \frac{A_c}{2} \cdot 2\boldsymbol{M} = \boldsymbol{M},\end{aligned} \tag{14}$$

which aligns with the standard attention mechanism outlined in Equation (2). This strategy ensures that the model initiates training with a well-established mechanism before incorporating more sophisticated modifications. Moreover, other modules, such as the normalization and FFN layers, are replicated directly from the original transformer block to ensure structural coherence.

3 Experiments

3.1 Training Settings

Training Data. We incorporate some noisy context data into the general supervised fine-tuning dataset to enhance LLMs' denoising capability in noisy context scenarios. This training involved integrating three distinct datasets: LongCite-45k (Zhang et al., 2024), Neural-Bridge-RAG (Neural Bridge AI, 2024) and Tulu3-SFT-Mix (Lambert et al., 2024). After data processing, we get the **N**oisy **C**ontext **F**ine-**T**uning (NCFT) dataset for supervised fine-tuning. We provide more details of the NCFT dataset in Appendix B.

OpAmp Models. We select two pre-trained models with different model sizes, Qwen2.5-72B (Yang et al., 2024) and Llama3.1-8B (Dubey et al., 2024), as our base models to train our OpAmp models using the NCFT dataset. Moreover, we use the QLoRA technique to update the other parameters in the pre-trained models instead of full fine-tuning. Please refer to Appendix A for more details.

3.2 Evalutaion Settings

Baselines. We compare our OpAmp models with existing powerful LLMs in our evaluation benchmark. These LLMs include Llama3-ChatQA2-70B (Xu et al., 2024), Qwen2.5-72B-inst (Yang et al., 2024), Llama3.3-70B-inst (Dubey et al., 2024), DeepSeek-V3 (Liu et al., 2024a), GPT-4o-0806 (Hurst et al., 2024), Llama3-ChatQA2-8B (Xu et al., 2024), Mistral-7B-inst-v0.3 (Jiang et al., 2023), Llama3.1-8B-inst (Meta, 2024) and Qwen2.5-7B-inst (Yang et al., 2024).

	Qwen2.5 OpAmp-72B	Llama3 ChatQA2-70B	Qwen2.5 72B inst	Llama3.3 70B inst	DeepSeek V3	GPT-4o 0806
LooGLE (EM) (Li et al., 2023)	**66.3**	59.1	64.9	63.0	63.4	62.7
NarrativeQA (EM) (Kočiskỳ et al., 2018)	**61.7**	59.8	60.2	61.5	60.5	61.5
MultiHopRAG (EM) (Tang and Yang, 2024)	**89.6**	78.2	89.2	83.7	88.6	87.7
HotpotQA (EM) (Yang et al., 2018)	**77.5**	70.5	76.0	74.5	77.0	**77.5**
MuSiQue (EM) (Trivedi et al., 2022)	48.0	39.0	44.0	47.5	52.5	**53.0**
CoQA (EM) (Reddy et al., 2019)	**92.4**	80.2	85.8	88.2	88.4	88.6

Table 1: Performance of Qwen2.5-OpAmp-72B on various noisy context benchmarks. We present a detailed comparison of the Qwen2.5-OpAmp-72B with current SOTA open-source and commercial LLMs. We bold the highest scores among all models.

	Llama3.1 OpAmp-8B	Llama3 ChatQA2-8B	Mistral 7B inst-v0.3	Llama3.1 8B inst	Qwen2.5 7B inst
LooGLE (EM)	**56.6**	50.7	51.6	56.1	53.8
NarrativeQA (EM)	**57.4**	53.1	44.7	55.9	47.7
MultiHopRAG (EM)	**70.5**	50.9	69.5	63.9	66.9
HotpotQA (EM)	**61.0**	56.5	58.0	58.5	59.5
MuSiQue (EM)	**35.0**	23.0	28.5	29.5	31.5
CoQA (EM)	**85.4**	78.2	70.6	82.2	84.2

Table 2: Performance of Llama3.1-OpAmp-8B on various noisy context benchmarks. We present a detailed comparison of the Llama3.1-OpAmp-8B with various open-source LLMs with similar parameters. We bold the highest scores among all models.

Evalution Benchmarks. Our evaluation benchmarks are designed using a spectrum of well-known datasets and benchmarks including LongBench (Yushi et al., 2024) and ChatQA (Liu et al., 2024c). After some selection and filtration, these benchmarks can be categorized as follows:

- **Long-Context QA:** The evaluation encompasses partial match (PM), exact match (EM), and accuracy (Acc.) metrics for various long-context QA benchmarks, including NarrativeQA (Kočiskỳ et al., 2018), Qasper (Dasigi et al., 2021), QuALITY (Pang et al., 2021), and LooGLE (Li et al., 2023).
- **Multi-Hop QA:** Assessment of multi-hop reasoning performance on various benchmarks, including HotpotQA (Yang et al., 2018), MuSiQue (Trivedi et al., 2022), and MultiHopRAG (Tang and Yang, 2024), using the EM metric.
- **Noisy-RAG QA:** PM and EM scores for RAG scenarios using CoQA (Reddy et al., 2019), QuAC (Choi et al., 2018), and QReCC (Anantha et al., 2020) benchmarks.

For a more detailed composition of the evaluation benchmark, please refer to Appendix C.

3.3 Evaluation on Noisy-Context Benchmarks

We perform various experiments on LLMs with different sizes to evaluate the capabilities of our OpAmp adaptation. For LLMs with more than 70B parameters, we compare Qwen2.5-OpAmp-72B with Llama3-ChatQA2-70B, Qwen2.5-72B-inst, Llama3.3-70B-inst, DeepSeek-V3, and GPT-4o-0806. For LLMs with around 7B parameters, we compare Llama3.1-OpAmp-8B with Llama3-ChatQA2-8B, Mistral-7B-inst-v0.3, Llama3.1-8B-inst, and Qwen2.5-7B-inst. The noisy-context benchmarks cover a wide range of tasks. For long-context scenarios, LooGLE and NarrativeQA are selected. We utilize MultiHopRAG, HotpotQA, and MuSiQue for Multi-Hop reasoning evaluation and CoQA for noisy-RAG scenarios. Table 1 and Table 2 demonstrate the superior performance of our OpAmp models compared to other existing powerful LLMs, underscoring the significant capabilities and effectiveness of the OpAmp adaptation

Method	$\mathcal{K}$	Avg.	Qasper (PM)	LooGLE (EM)	NarrativeQA (EM)	QuALITY (Acc.)	MultiHopRAG (EM)	HotpotQA (EM)	MuSiQue (EM)	CoQA (EM)	QuAC (PM)	QReCC (PM)
QLoRA	-	52.4	38.9	53.1	55.7	76.1	68.4	56.5	31.5	83.6	25.2	35.4
OpAmp Adapter	1	54.1 (+1.7)	40.8	56.0	56.4	79.2	68.5	57.5	32.5	**85.8**	26.1	38.3
	5	54.3 (+1.9)	41.2	56.5	56.9	77.8	69.5	**62.0**	31.5	84.6	25.5	37.1
	10	**55.4 (+3.0)**	**43.1**	**56.6**	**57.4**	79.0	70.5	61.0	**35.0**	85.4	**26.5**	**39.8**
	20	54.4 (+2.0)	41.5	55.4	56.4	**79.3**	**71.4**	59.0	33.0	84.0	26.2	37.0

Table 3: Ablation studies on various noisy context benchmarks using Llama3.1-8B-base as the base model. We bold the highest scores for each benchmark.

in noisy context scenarios.

Long-Context Evaluation. Long-context evaluation requires LLMs to disregard large volumes of context-related but question-irrelevant information within extensive texts, accurately identify the paragraphs relevant to the answer, and generate responses based on these pertinent segments. Our Qwen2.5-OpAmp-72B model achieves EM scores of up to 66.3% on the LooGLE benchmark with a maximum context length of 32K tokens and 61.7% on the NarrativeQA benchmark with a maximum context length of 64K tokens. Similarly, our Llama3.1-OpAmp-8B model attains the highest EM score of 56.6% on the LooGLE benchmark and leads with a score of 57.4% on the NarrativeQA benchmark. These experiment results underscore the robust capability of our OpAmp models to filter out context-related noise and accurately locate answers within long contexts. Furthermore, they demonstrate the strong generalization ability of our approach across different model sizes.

Multi-Hop Evaluation. Multi-hop evaluation is designed to assess the capability of LLMs to extract and synthesize relevant information from multiple documents for reasoning. This task requires LLMs to filter out irrelevant or noncritical documents to minimize interference during the reasoning process. Our Qwen2.5-OpAmp-72B model demonstrates strong performance on multi-hop reasoning tasks, achieving high scores of 89.6% on MultiHopRAG and 77.5% on HotpotQA, with notable advantages over competing models. Although it performs slightly weaker than top-performing LLMs on the MuSiQue benchmark, its EM score of 48.0% remains competitive for multi-hop reasoning tasks. Additionally, our Llama3.1-OpAmp-8B model also excels in multi-hop reasoning benchmarks, achieving top scores of 70.5% on MultiHopRAG, 61.0% on HotpotQA, and 35.0% on MuSiQue, consistently surpassing other models. These results highlight the superior ability of our OpAmp models to handle complex, multi-step reasoning tasks across various benchmarks, underscoring its effectiveness in enhancing reasoning capabilities.

Noisy-RAG Evaluation. For the currently most widely adopted RAG technology, we conduct the noisy-RAG evaluation to assess the ability of LLMs to filter out irrelevant documents and accurately identify the document containing the correct answer in real-world RAG scenarios. Our Qwen2.5-OpAmp-72B model achieves a top score of 92.4% on the CoQA benchmark, surpassing the second-closest LLM, DeepSeek-V3, by a significant margin of 4%. Our Llama3.1-OpAmp-8B model also attains a leading score of 85.4% on the CoQA benchmark, outperforming Qwen2.5-7B-inst by 1.2%. These experimental results highlight the superior performance of our OpAmp models in identifying correct answers within real-world RAG scenarios, exhibiting robust resistance to interference and noise.

3.4 Ablation Studies

To further investigate the contribution of $\mathcal{K}$, we perform a series of ablation studies. Additionally, we compare our OpAmp approach with the QLoRA (Dettmers et al., 2024) technique. In brief, we denote the OpAmp adapter as the adapter implemented for our OpAmp adaptation. The QLoRA technique performs PEFT without modifying the pre-trained model's attention mechanism. To ensure fair comparisons in these ablation studies, both OpAmp and QLoRA models are fine-tuned using the same dataset, NCFT.

CMRR. Table 3 presents a comparative analysis of QLoRA and the OpAmp adapter for enhancing the Llama3.1-8B-base model across various noisy context benchmarks. The OpAmp adapter demonstrates consistent superiority over QLoRA across all evaluated benchmarks. Specifically, QLoRA achieves an average score of 52.4%, whereas the OpAmp adapter significantly enhances perfor-

		CoQA (EM)			QuAC (PM)			QReCC (PM)		
Noise Ratio		0.0	0.8	0.9	0.0	0.8	0.9	0.0	0.8	0.9
QLoRA		89.8	85.4	83.6	27.5	26.1	25.2	36.5	36.4	35.4
OpAmp Adapter	$\mathcal{K} = 1$	90.4	85.6	**85.8**	28.5	26.2	26.1	39.4	39.1	38.3
	$\mathcal{K} = 5$	90.0	85.6	84.6	27.5	26.7	25.5	38.2	37.3	37.1
	$\mathcal{K} = 10$	91.2	**88.0**	85.4	28.5	26.5	**26.5**	**40.8**	**39.8**	**39.8**
	$\mathcal{K} = 20$	**91.8**	86.6	84.0	**28.6**	**28.0**	26.2	38.5	38.1	37.0

Table 4: Ablation studies on various benchmarks with different noise ratios using Llama3.1-8B-base as the base model. We bold the highest scores.

Method	$\mathcal{K}$	FaithEval			
		Inconsistent (EM)	Unanswerable (EM)	Counterfactual (EM)	Avg.
QLoRA	-	24.1	46.1	71.6	47.3
OpAmp Adapter	1	**45.5**	53.1	**76.3**	**58.3 (+11.0)**
	5	42.1	**53.7**	75.9	57.2 (+9.90)
	10	45.3	53.0	75.1	57.8 (+10.5)
	20	22.3	58.8	73.8	51.6 (+4.30)

Table 5: Ablation studies on FaithEval using Llama3.1-8B-base as the base model. We bold the highest scores.

mance, with the best results observed at $\mathcal{K} = 10$, yielding an average score of 55.4%. When examining the impact of different values of $\mathcal{K}$, $\mathcal{K} = 10$ emerges as the optimal configuration across multiple benchmarks. Larger value ($\mathcal{K} = 20$) exhibits diminishing returns, while smaller values ($\mathcal{K} = 1, 5$) perform adequately but are marginally less competitive. This suggests our statement that attention denoising requires only a modest $\mathcal{K}$ instead of the $\mathcal{K} \to \infty$ used in the differential transformer architecture (Ye et al., 2025).

Noise Ratio. The ablation study detailed in Table 4 assesses the performance of QLoRA and the OpAmp adapter across varying noise ratios (0.0, 0.8, 0.9) on noisy-RAG benchmarks, including CoQA, QuAC, and QReCC. The noise ratio is simulated by introducing noise documents into the original golden document, replicating increasingly challenging real-world RAG scenarios. As expected, performance across all methods generally degrades with increasing noise ratios, reflecting the growing difficulty of extracting relevant information from cluttered contexts. QLoRA exhibits a steady decline in performance as noise levels increase. For instance, its score on CoQA drops from 89.8% at a noise ratio of 0.0 to 83.6% at 0.9. In contrast, the OpAmp adapter demonstrates greater robustness, particularly when configured with $\mathcal{K} = 10$. Moreover, higher values of $\mathcal{K}$ occasionally underperform compared to $\mathcal{K} = 10$, indicating that excessive attention denoise may compromise the capability. Overall, the OpAmp adapter consistently outperforms QLoRA across all noise levels, with $\mathcal{K} = 10$ emerging as the optimal configuration for balancing robustness and performance under noisy conditions. This underscores the effectiveness of our method in handling challenging RAG scenarios.

Hallucination. As shown in Table 5, the ablation study on FaithEval (Ming et al., 2024) demonstrates that OpAmp not only enhances robustness to noisy contexts but also reduces hallucinations as a valuable secondary benefit. While QLoRA achieves an average score of 47.3%, OpAmp attains much higher averages, with the best results with $\mathcal{K} = 1$ (58.3%), indicating consistent improvements. Notably, $\mathcal{K} = 1, 5, 10$ exhibit similar performance levels, suggesting that moderate values of $\mathcal{K}$ effectively balance denoising and model stability while mitigating hallucinations. However, performance declines significantly (51.6%) when $\mathcal{K} = 20$. The degradation demonstrates an excessive attention-denoising process caused by excessive CMRR, which impairs the model's ability to avoid hallucination. This analysis underscores that the optimal performance is achieved with moderate $\mathcal{K}$ values, highlighting the importance of balancing denoising intensity with model adaptability.

3.5 Visualization of Attention

To provide deeper insights into the OpAmp mechanism, we perform some visualizations of $\bar{\boldsymbol{M}}$. As previously mentioned, transformer-based architec-

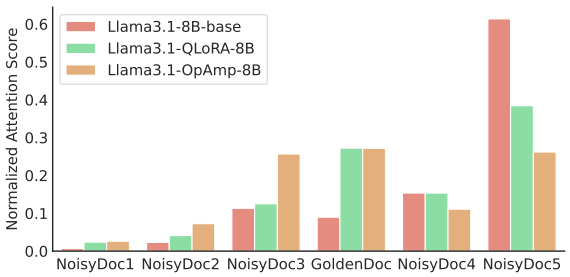

Figure 5: Normalized attention score. Our OpAmp model demonstrates significant attention denoise capability compared to the base model and QLoRA model.

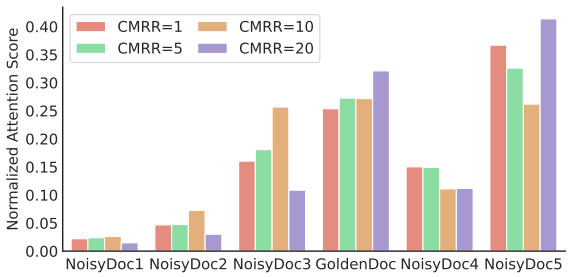

Figure 6: Normalized attention score with different values of $\mathcal{K}$ utilizing for OpAmp adaptation.

tures tend to allocate disproportionate attention to irrelevant or later-positioned documents. In contrast, OpAmp can enhance LLMs' focus on the most relevant documents. We employ normalized attention scores based on Llama3.1-8B to trace the OpAmp mechanism in a noisy context to investigate this behavior. As shown in Figure 5, Llama3.1-8B-base becomes completely lost in the noisy context, with its attention distribution across documents generally increasing sequentially from low to high. Llama3.1-QLoRA-8B model performs relatively better, with a slight increase in attention to the golden document. However, the limitation of a forced backward shift in attention still exists. In contrast, our Llama3.1-OpAmp-8B uniquely allocates the most attention to the golden document among all documents. This mechanism is a key factor contributing to the strong performance of our OpAmp model in noisy context scenarios. Meanwhile, we also investigate the mechanism across different CMRR values. As illustrated in Figure 6, only when $\mathcal{K} = 10$ does the OpAmp model allocate the highest level of attention to the golden document, surpassing the other CMRR values and indirectly confirming that a moderate CMRR value is crucial for maximizing the effectiveness of the OpAmp mechanism instead of $\mathcal{K} \to \infty$ utilized in differential transformer (Ye et al., 2025).

4 Discussion

In this paper, we primarily investigate how our architectural adaptation influences attention distributions and model performance.

We acknowledge the need for a practical solution to mitigate overhead, and here are some theoretically viable approaches. Firstly, we can store the post-adaptation K cache during decoding to avoid recomputation. Secondly, we can halve the number of groups in GQA (Ainslie et al., 2023), thereby preserving computational efficiency and ensuring no additional memory overhead. This mirrors techniques in the Differential Transformer (Ye et al., 2025), where reducing the number of heads (while maintaining or surpassing the performance of models with $2\times$ heads) keeps memory costs stable.

Regarding the functional differences between the two distinct attention score matrices, V_{in}^{+} and V_{in}^{-}, we observe that their roles are highly context-dependent, varying across different data and model layers. Nevertheless, the coupled mechanism effectively enhances the model's focus on the most relevant (golden) document, as illustrated in Section 3. We acknowledge the complexity of this phenomenon and plan to investigate the precise roles of V_{in}^{+} and V_{in}^{-} in future work.

5 Related Works

5.1 Question Answering with Noisy Contexts

The internal knowledge of LLMs often fails to meet diverse application needs (He et al., 2022; Ji et al., 2023), driving research into integrating external knowledge. Among the proposed solutions (Guu et al., 2020; Beltagy et al., 2020; Wang et al., 2024), RAG (Borgeaud et al., 2022; Ren et al., 2024) and long-context modeling techniques (Press et al., 2022; Chen et al., 2023b) have emerged as two prominent strategies for incorporating external knowledge stored in long-text formats. However, recent studies (Shi et al., 2023; Liu et al., 2024b; Lv et al., 2024; Ye et al., 2025) have identified a significant challenge. Specifically, as the number of retrieved documents grows and the length of input contexts expands, the model is increasingly exposed to noise, which is often the non-critical information unrelated to the query. This noisy-context scenario significantly degrades the performance of LLMs on QA tasks (Chen et al., 2023a). Consequently, we propose the OpAmp adaptation with

adapters, a plug-and-play solution that minimizes noisy context impact with low computation costs, enhancing the performance in such scenarios.

5.2 Parameter Efficient Fine-Tuning

Traditionally, full fine-tuning is the predominant approach for fine-tuning pre-trained models, including LLMs. However, this method entails substantial computational costs, particularly regarding time consumption and GPU memory usage. To address these challenges, a variety of PEFT methods have been developed (Houlsby et al., 2019; Hu et al., 2021; Dettmers et al., 2024; Wu et al., 2024b; Li and Liang, 2021; Lester et al., 2021; Wu et al., 2024a), enabling efficient fine-tuning without compromising performance compared to full fine-tuning. PEFT focuses on training a limited subset of parameters within the existing model or newly inserted modules. Adapter-based methods (Houlsby et al., 2019; Hu et al., 2021; Dettmers et al., 2024; Wu et al., 2024b) insert learnable modules into Transformer blocks, which contain a small number of parameters. These adapters are fine-tuned instead of the original model weights. Among these methods, QLoRA (Dettmers et al., 2024) has gained significant attention for its efficiency in fine-tuning LLMs while maintaining performance comparable to full fine-tuning. Another emerging trend in PEFT is prefix-tuning (Lester et al., 2021; Li and Liang, 2021), which involves adding learnable token vectors to the input sequence. In this study, we introduce adapters to perform OpAmp adaptation. Specifically, adapters reformulate the computation of the original attention mechanism into the OpAmp attention mechanism.

5.3 Adaptation of Pre-trained Models

Recent studies (Chen et al., 2015; Lin et al., 2021; Komatsuzaki et al., 2023; Wu et al., 2024b) have focused on improving training efficiency by leveraging pre-trained model weights for a warm start, thus accelerating convergence and minimizing training costs. Komatsuzaki et al. (2023) and Wu et al. (2024b) introduce methods to initialize sparse MoE models using weights from a pre-trained dense model. These approaches significantly reduce the required training resources. In this paper, we train our OpAmp models with OpAmp attention blocks using weights from pre-trained LLMs.

6 Conclusion

Inspired by the operational amplifiers, we introduce the OpAmp adaptation implemented with adapters in this study. By integrating this adapter into pre-trained Transformer blocks, our approach enhances the model's ability to focus on the most relevant context without expensive full-scale training from scratch. We implement our OpAmp models and other baselines with our noisy-context fine-tuning dataset, NCFT, for fair comparisons. The OpAmp adaptation demonstrates significant performance gains across LLMs of varying model sizes. Extensive empirical evaluations are conducted on extensive noisy-context benchmarks. The results indicate that our Qwen2.5-OpAmp-72B model, fine-tuned with our OpAmp adaptation, outperforms current SOTA LLMs, including DeepSeek-V3 (Liu et al., 2024a) and GPT-4o (Hurst et al., 2024).

Limitation

The OpAmp adaptation with adapters introduces a marginally higher number of parameters compared to the standard PEFT training process with QLoRA. Consequently, the supervised fine-tuning process for our OpAmp models demands slightly greater GPU memory allocation and computational time. Additionally, our OpAmp models incur a minor latency during inference when compared to the original pre-trained LLMs.

Acknowledgement

This work is partially supported by The Research Grants Council of Hong Kong SAR (No. RFS2425-4S02, CUHK14211824, and CUHK14201624), and the MIND project (MINDXZ202404).

References

Joshua Ainslie, James Lee-Thorp, Michiel De Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. GQA: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2020. Open-domain question answering goes conversational via question rewriting. *arXiv preprint arXiv:2010.04898*.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. *arXiv preprint arXiv:2112.04426*.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2023a. Benchmarking large language models in retrieval-augmented generation. *arXiv preprint arXiv:2309.01431*.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023b. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

Tianqi Chen, Ian Goodfellow, and Jonathon Shlens. 2015. Net2Net: Accelerating learning via knowledge transfer. *arXiv preprint arXiv:1511.05641*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. *arXiv preprint arXiv:1808.07036*.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. 2021. A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers. *arXiv preprint arXiv:2105.03011*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. QLoRA: Efficient finetuning of quantized LLMs. In *Annual Conference on Neural Information Processing Systems (NIPS)*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Hangfeng He, Hongming Zhang, and Dan Roth. 2022. Rethinking with retrieval: Faithful large language model inference. *arXiv preprint arXiv:2301.00303*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. GPT-4o System Card. *arXiv preprint arXiv:2410.21276*.

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*.

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

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA Reading Comprehension Challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. 2023. Sparse Upcycling: Training mixture-of-experts from dense checkpoints. In *International Conference on Learning Representations (ICLR)*.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. 2024. Tülu 3: Pushing Frontiers in Open Language Model Post-Training. *arXiv preprint arXiv:2411.15124*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2023. LooGLE: Can Long-Context Language Models Understand Long Contexts? *arXiv preprint arXiv:2311.04939*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Junyang Lin, An Yang, Jinze Bai, Chang Zhou, Le Jiang, Xianyan Jia, Ang Wang, Jie Zhang, Yong Li, Wei Lin, et al. 2021. M6-10T: A sharing-delinking paradigm for efficient multi-trillion parameter pretraining. *arXiv preprint arXiv:2110.03888*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. DeepSeek-V3 Technical Report. *arXiv preprint arXiv:2412.19437*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024b. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*.

Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. 2024c. Chatqa: Surpassing gpt-4 on conversational qa and rag. *arXiv preprint arXiv:2401.10225*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled Weight Decay Regularization. *arXiv preprint arXiv:1711.05101*.

Qitan Lv, Jie Wang, Hanzhu Chen, Bin Li, Yongdong Zhang, and Feng Wu. 2024. Coarse-to-Fine Highlighting: Reducing Knowledge Hallucination in Large Language Models. *arXiv preprint arXiv:2410.15116*.

AI Meta. 2024. Introducing Llama 3.1: Our most capable models to date. *Meta AI Blog*.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2024. FaithEval: Can Your Language Model Stay Faithful to Context, Even If" The Moon is Made of Marshmallows". *arXiv preprint arXiv:2410.03727*.

Neural Bridge AI. 2024. Retrieval-Augmented Generation (RAG) Dataset 12000.

OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, et al. 2021. QuALITY: Question Answering with Long Input Texts, Yes! *arXiv preprint arXiv:2112.08608*.

Ofir Press, Noah A. Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2024. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *arXiv preprint arXiv:2307.11019*.

Willy M Sansen. 2007. *Analog design essentials*, volume 859. Springer Science & Business Media.

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

Yixuan Tang and Yi Yang. 2024. MultiHop-RAG: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv preprint arXiv:2401.15391*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop Questions via Single-hop Question Composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2024. Knowledge editing for large language models: A survey. *arXiv preprint arXiv:2310.16218*.

Haoyuan Wu, Xinyun Zhang, Peng Xu, Peiyu Liao, Xufeng Yao, and Bei Yu. 2024a. p-Laplacian adaptation for generative pre-trained vision-language models. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Haoyuan Wu, Haisheng Zheng, Zhuolun He, and Bei Yu. 2024b. Parameter-Efficient Sparsity Crafting from Dense to Mixture-of-Experts for Instruction Tuning on General Tasks. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Peng Xu, Wei Ping, Xianchao Wu, Chejian Xu, Zihan Liu, Mohammad Shoeybi, and Bryan Catanzaro. 2024. ChatQA 2: Bridging the gap to proprietary LLMs in long context and RAG capabilities. *arXiv preprint arXiv:2407.14482*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv: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. *arXiv preprint arXiv:1809.09600*.

Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. 2025. Differential Transformer. In *International Conference on Learning Representations (ICLR)*.

Bai Yushi, Lv Xin, Zhang Jiajie, Lyu Hongchang, Tang Jiankai, Huang Zhidian, Du Zhengxiao, Liu Xiao, Zeng Aohan, Hou Lei, Dong Yuxiao, Tang Jie, and Li Juanzi. 2024. 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)*.

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. 2024. LongCite: Enabling LLMs to Generate Fine-grained Citations in Long-context QA. *arXiv preprint arXiv:2409.02897*.

lr	epoch	LoRA r	LoRA α	Adapter Dim
2×10^{-4}	1	64	16	512

Table 6: Hyperparameters of supervised fine-tuning.

	LongCite-45k	Neural-Bridge-RAG	Tulu3-SFT-Mix
NCFT	30k	20k	450k

Table 7: The proportion of LongCite-45k, Neural-Bridge-RAG and Tulu3-SFT-Mix in the NCFT dataset.

A Implementation Details

The training process entailed using a constant learning rate schedule with a warm-up ratio of 0.03, and the paged AdamW (Dettmers et al., 2024; Loshchilov and Hutter, 2017) optimizer with a learning rate of 2×10^{-4}, no weight decay, a batch size of 128, and a sequence length of 8192 tokens. The models underwent instruction tuning for one epoch on 16 A100 GPUs, each with 80G memory.

Moreover, we employed the QLoRA (Dettmers et al., 2024) technique for efficient fine-tuning. As for the QLoRA configuration, we use a 4-bit quantization scheme for our experiments, which significantly reduces memory usage while preserving model performance. We show the hyperparameters for supervised fine-tuning in Table 6.

B Training Datasets

As shown in Table 7, we shows the proportion of LongCite-45k (Zhang et al., 2024), Neural-Bridge-RAG (Neural Bridge AI, 2024) and Tulu3-SFT-Mix (Lambert et al., 2024) in the NCFT dataset.

Considering the original format and quantity of LongCite-45k and Neural-Bridge-RAG, we perform data processing to simulate the noisy context scenarios. Firstly, we filter the Chinese corpus and divide the context into several chunks. Then we preserve the chunks with golden documents and introduce relevant or irrelevant chunks as noise. Finally, we filter low-quality corpora (too long or too short). We obtained our supervised fine-tuning dataset after data processing which encompasses a wide range of topics, and the noise ratio in the dataset ranges from 0 to 1, aiming to cover a variety of real-world situations and use cases.

C Evaluation Benchmarks

We show the details of the noisy-context evaluation benchmark in Table 8. Qasper, HotpotQA, and MuSiQue are directly derived from the LongBench (Yushi et al., 2024). In contrast, CoQA, QuAC, and QReCC are QA datasets selected from ChatQA (Liu et al., 2024c) and have been noise-augmented in a manner consistent with Appendix B to align with the noisy-RAG format. For the QuALITY dataset, we retain only the subset labeled as "hard". Similarly, for the NarrativeQA, Loogle, and MultiHopRAG datasets, we apply filters based on context length and response quality to further enhance the benchmark's ability to differentiate between models.

Benchmark	Source	Max Length	Metric	# Data
Long-Context QA				
NarrativeQA	Literature, Film	64K	EM	1009
Qasper	Science	8K	PM	200
QuALITY	Literature	8K	Acc.	1065
LooGLE	Science	32K	EM	1427
Multi-Hop QA				
HotpotQA	Wikipedia	16K	EM	200
MuSiQue	Wikipedia	16K	EM	200
MultiHopRAG	News	8K	EM	2255
Noisy-RAG QA				
CoQA	Multi-field	4K	EM	500
QuAC	Wikipedia	4K	PM	996
QReCC	Multi-field	4K	PM	643

Table 8: An overview of the dataset statistics for the noisy-context benchmark. The 'Source' column indicates the origin of the context.

The curation of datasets is primarily based on two criteria to ensure benchmark quality and fairness. Firstly, we removed samples with excessively short contexts to mitigate noise and maintain high-quality evaluation. Secondly, since Exact Match (EM) is our primary metric, we exclude overly long ground-truth answers to ensure EM's evaluation stability. Shorter ground truths align better with EM's design, enabling reliable model performance comparisons. These filtering steps are model-agnostic and do not favor any specific approach, thus preserving the benchmark's impartiality.