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

ACL 2025 VIENNA

Introduction

Background:

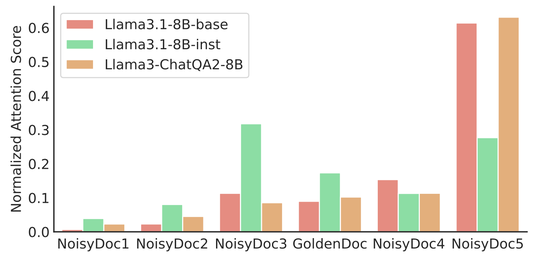

Figure 1. Normalized attention score. Transformers often miss the golden document in a noisy context.

Contribution:

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. 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.

Operational Amplifier

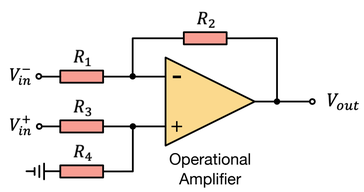

Figure 2. The operational amplifier with two input voltages V_{in}^+ and V_{in}^-. The CMRR $\mathcal{K}$ is controlled by resistances R_1, R_2, R_3, R_4.

$$V_{\text{out}} = V_{\text{in}}^+ \cdot \left(\frac{R_4}{R_3+R_4} \cdot \frac{R_1+R_2}{R_1}\right) - V_{\text{in}}^- \cdot \frac{R_2}{R_1}$$
$$= A_d(V_{\text{in}}^+ - V_{\text{in}}^-) + \frac{A_c}{2}(V_{\text{in}}^+ + V_{\text{in}}^-). \quad (1)$$

OpAmp Adaptation

Inspired by the operational amplifier, we propose the OpAmp adaptation, which modifies the original attention mechanism into the OpAmp attention mechanism.

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

where $\vec{\vec{M}}$ is the denoised attention matrix via OpAmp adaptation, $\vec{M}^+$ and $\vec{M}^-$ are formulated through adapters

Architecture

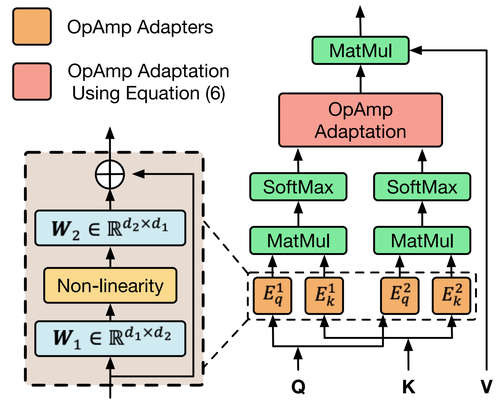

Figure 3. Overview of the OpAmp adaptation with adapters.

Zero Initialization

At the onset of training, we employ zero initialization to promote identity mapping. Specifically, $\vec{W}_2$ is initialized to zero to guarantee that $E_j^i(\vec{x}) = \vec{x}$. Furthermore, to prevent any disruption to the original $\vec{M}$ 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 (2) reduces to:

$$\vec{\vec{M}} = A_d \cdot (\vec{M} - \vec{M}) + \frac{A_c}{2} \cdot (\vec{M} + \vec{M}),$$
$$= A_d \cdot 0 + \frac{A_c}{2} \cdot 2\vec{M} = \vec{M}, \quad (3)$$

Experiment Settings

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

Table 1. The proportion of LongCite-45k, Neural-Bridge-RAG and Tulu3-SFT-Mix in our training dataset.

Experiment Results

	Qwen2.5 OpAmp-72B	Llama3 ChatQA2-70B	Qwen2.5 72B inst	Llama3.3 70B inst	DeepSeek V3	GPT-4o 0806
LooGLE	66.3	59.1	64.9	63.0	63.4	62.7
NarrativeQA	61.7	59.8	60.2	61.5	60.5	61.5
MultiHopRAG	89.6	78.2	89.2	83.7	88.6	87.7
HotpotQA	77.5	70.5	76.0	74.5	77.0	77.5
MuSiQue	48.0	39.0	44.0	47.5	52.5	53.0
CoQA	92.4	80.2	85.8	88.2	88.4	88.6

Table 2. 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.

Experiment Results

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

Table 3. 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.

Hallucination

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

Visualization of Attention

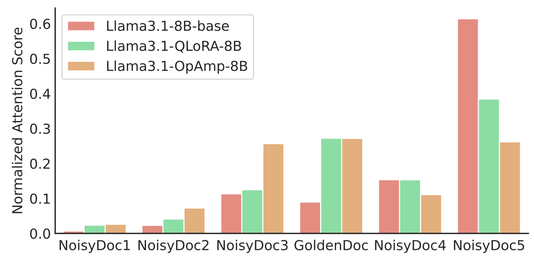

Figure 4. Normalized attention score. Our OpAmp model demonstrates significant attention denoise capability compared to the base model and QLoRA model.

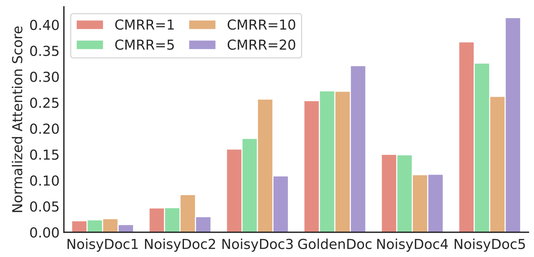

Figure 5. Normalized attention score with different values of $\mathcal{K}$ utilizing for OpAmp adaptation