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
[♡] Shanghai Artificial Intelligent Laboratory
[♣] ChatEDA Technology

Jul. 27, 2025

Outline

Introduction

Background

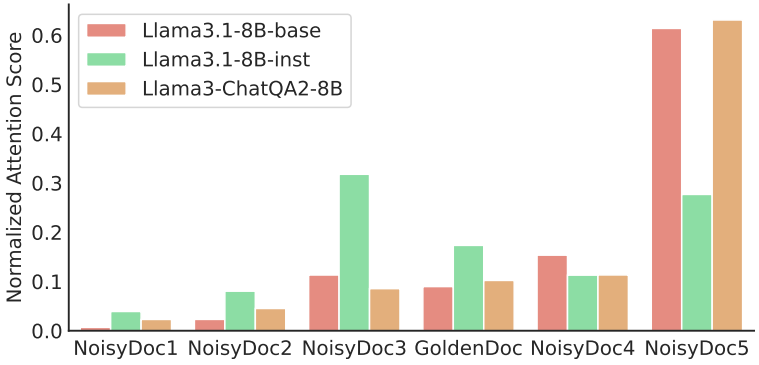

Figure: Normalized Attention Score.

- Transformers often miss the golden document in a noisy context.

Contribution

- 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.

Method

Operational Amplifier

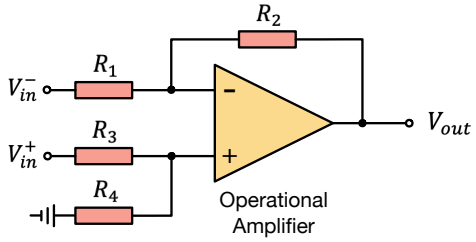

Figure: Visualization of *Operational Amplifier*.

- 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.

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

Common-Mode Rejection Ratio

The *Common-Mode Rejection Ratio (CMRR)* is defined as the ratio of the differential gain to the common-mode gain:

$$\mathcal{K} = \frac{A_d}{A_c}. \tag{2}$$

- *CMRR* is a very important parameter, which indicates the suppression and attenuation of the common-mode signal passing through the amplifier.

OpAmp Adaptation

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

$$\bar{\boldsymbol{M}} = A_d(\boldsymbol{M}^+ - \boldsymbol{M}^-) + \frac{A_c}{2}(\boldsymbol{M}^+ + \boldsymbol{M}^-), \tag{3}$$

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

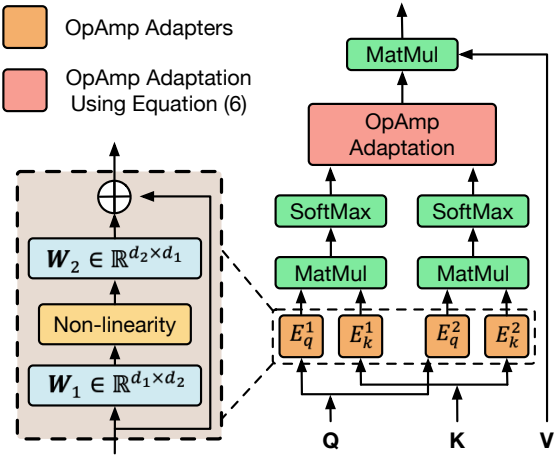

Figure: Overview of the OpAmp adaptation using Equation (3) with adapters.

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 (3) reduces to:

$$\begin{aligned} \bar{\boldsymbol{M}} &= A_d \cdot (\boldsymbol{M} - \boldsymbol{M}) + \frac{A_c}{2} \cdot (\boldsymbol{M} + \boldsymbol{M}), \\ &= A_d \cdot 0 + \frac{A_c}{2} \cdot 2\boldsymbol{M} = \boldsymbol{M}, \end{aligned} \tag{4}$$

Experiments

Experiments Setting

	LongCite-45k [1]	Neural-Bridge-RAG [2]	Tulu3-SFT-Mix [3]
NCFT	30k	20k	450k

Table: Training Dataset Composition and Proportions.

[1] **Jiajie Zhang et al. (2024). "LongCite: Enabling LLMs to Generate Fine-grained Citations in Long-context QA".** In: *arXiv preprint*.

[2] **Neural Bridge AI (2024).** ***Retrieval-Augmented Generation (RAG) Dataset 12000.*** URL: `https://huggingface.co/datasets/neural-bridge/rag-dataset-12000`.

[3] **Nathan Lambert et al. (2024). "Tülu 3: Pushing Frontiers in Open Language Model Post-Training".** In: *arXiv preprint*.

Experiments Setting

ACL 2025 VIENNA

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: An Overview of the Dataset Statistics for the *Noisy-Context* Benchmark.

Evaluation on Noisy-Context Benchmarks

ACL 2025 VIENNA

	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: Performance of Qwen2.5-OpAmp-72B on various *Noisy Context Benchmarks*.

Evaluation on Noisy-Context Benchmarks

ACL 2025 VIENNA

	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
MultiHopRAG	**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: Performance of Llama3.1-OpAmp-8B on various *Noisy Context Benchmarks*.

Ablation Studies: Common-Mode Rejection Ratio

ACL 2025 VIENNA

Method	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: Ablation Studies on various *Noisy Context Benchmarks* using Llama3.1-8B-Base as the Base Model.

Ablation Studies: Noise Ratios

		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: Ablation Studies on *Different Noise Ratios* using Llama3.1-8B-Base as the Base Model.

Ablation Studies: 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: Ablation Studies on *FaithEval* using Llama3.1-8B-Base as the Base Model.

Visualization of Attention

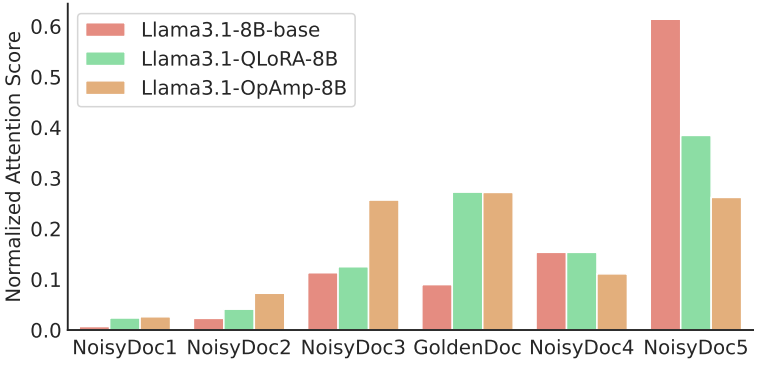

Figure: Normalized Attention Score.

- Our *OpAmp Model* demonstrates significant *Attention Denoise* capability compared to the base model and QLoRA model.

Visualization of Attention

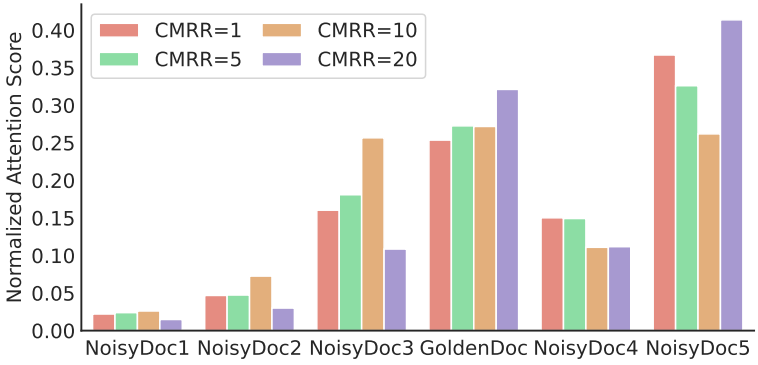

Figure: Normalized Attention Score with *Different Values of K Utilizing* for *OpAmp Adaptation*.

THANK YOU!