δ-mem: Efficient Online Memory for Large Language Models

Jingdi Lei[1], Di Zhang[2], Junxian Li[3], Weida Wang[2], Kaixuan Fan[4], Xiang Liu[5], Qihan Liu[6], Xiaoteng Ma[6], Baian Chen[6], Soujanya Poria[1]

[1]Nanyang Technological University, [2]Fudan University, [3]Shanghai Jiao Tong University,
[4]The Chinese University of Hong Kong, [5]Hong Kong University of Science and Technology (Guangzhou), [6]Mind Lab

Code Available

Motivation

Long-horizon assistants and agents must **accumulate and reuse history**, but:

- Longer context is **costly**: attention cost grows quadratically with length.
- Longer context ≠ better use: models suffer **context rot** on very long inputs.
- Existing memory mechanisms each fall short:

Memory type	Stores memory as	Limitation
Textual (TMM)	text in the prompt (RAG, LLMLingua-2, MemoryBank)	context budget, retrieval noise, compaction loss
Outside-channel (OMM)	external module + retrieval (MLP Memory)	extra overhead, integration complexity
Parametric (PMM)	prefixes / adapters (Context2LoRA)	static: adapts poorly to evolving information

Can a compact, continuously updated memory state steer a frozen LLM through its own attention?

Our Contributions

- **δ-mem**: an online associative memory for a **frozen** full-attention LLM; history is compressed into a fixed **8×8 state** written with **delta-rule learning**.
- The state readout yields **low-rank corrections** to the attention query and output: no full fine-tuning, no backbone replacement, no context extension.
- **1.10×** the frozen backbone and **1.15×** the strongest non-δ-mem memory baseline on average; **1.31×** on MemoryAgentBench and **1.20×** on LoCoMo.

Key Observations

- **Delta-rule writes correct memory via residuals**: associations that are already learned cause negligible updates.
- **Per-dimension gates** let some dimensions retain memory while others write actively.
- **An 8×8 state keeps usable history**: with explicit context removed, δ-mem still recovers part of the missing evidence.

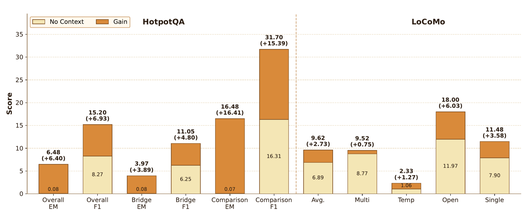


Figure 2: Context recovery on HotpotQA and LoCoMo (Qwen3-4B-Instruct) when explicit context is unavailable.

δ-mem: Delta-rule Online Memory

A frozen backbone plus a tiny online state $\mathbf{S}_t \in \mathbb{R}^{r\times r}$ ($r = 8$) that is **read** before and **written** after every attention step.

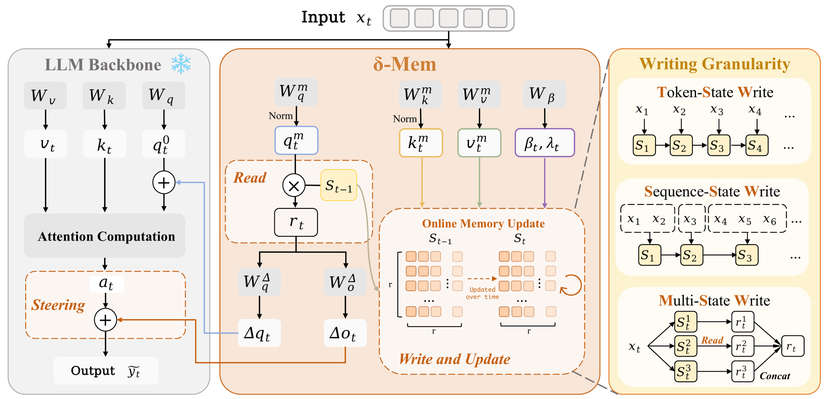


Figure 1: Overview of δ-mem.

1 Project the hidden state $\mathbf{x}_t$ into an r-dim memory space, with per-dimension gates:

$$\mathbf{q}_t^m,\ \mathbf{k}_t^m = \mathrm{L2norm}\left(\tanh\left(\mathbf{W}_{q,k}^m\,\mathbf{x}_t\right)\right), \qquad \mathbf{v}_t^m = \mathbf{W}_v^m\,\mathbf{x}_t$$
$$\boldsymbol{\beta}_t = \sigma\left(\mathbf{W}_\beta\,\mathbf{x}_t + \mathbf{b}\right), \qquad \boldsymbol{\lambda}_t = 1 - \boldsymbol{\beta}_t$$

2 **Read** from the previous state (cost independent of history length):

$$\mathbf{r}_t = \mathbf{S}_{t-1}\,\mathbf{q}_t^m$$

3 Apply **low-rank corrections** to the attention query and output:

$$\tilde{\mathbf{q}}_t = \mathbf{W}_Q\,\mathbf{x}_t + \frac{\alpha}{r}\,\mathbf{W}_q^\Delta\,\mathbf{r}_t$$
$$\tilde{\mathbf{y}}_t = \mathrm{Attn}(\tilde{\mathbf{q}}_t, \mathbf{K}_{\le t}, \mathbf{V}_{\le t}) + \frac{\alpha}{r}\,\mathbf{W}_o^\Delta\,\mathbf{r}_t$$

4 **Gated delta-rule write** of the new key–value pair:

$$\mathbf{S}_t = \underbrace{\mathrm{Diag}(\boldsymbol{\lambda}_t)\,\mathbf{S}_{t-1}}_{\text{retain}} + \mathrm{Diag}(\boldsymbol{\beta}_t)\,\underbrace{(\mathbf{v}_t^m - \mathbf{S}_{t-1}\mathbf{k}_t^m)}_{\text{prediction error}}\,(\mathbf{k}_t^m)^\top$$

Training. The context C is written into $\mathbf{S}_C$ but **not replayed** as input; only the δ-mem parameters θ are trained, the backbone φ stays frozen:

$$\mathcal{L}_{\mathrm{SFT}} = -\sum_{j=1}^{|Y|} \log p_{\phi,\theta}\left(y_j \mid Q,\, y_{<j},\, \mathbf{S}_C\right)$$

Writing strategies

TSW
Token-State Write: one write per token

SSW
Sequence-State Write: mean-pool each message, one write per message

MSW
Multi-State Write: N = 4 parallel sub-states, readouts concatenated

Main Results

Qwen3-4B-Instruct backbone. MAB = MemoryAgentBench; Avg = final average.

Method	IFEval	HotpotQA F1	GPQA-D	MAB	LoCoMo	Avg
Qwen3-4B-Instruct	81.89	56.00	39.39	29.54	40.79	46.79
+ BM25 RAG	–	52.83	–	24.49	36.68	44.56
+ LLMLingua-2	–	50.03	–	15.63	40.98	42.96
+ MemoryBank	–	–	–	17.65	38.14	43.88
+ Context2LoRA	76.71	50.88	29.29	32.53	48.11	44.90
+ MemGen	39.37	16.27	38.89	29.61	40.05	30.66
+ MLP Memory	24.95	25.83	22.73	28.80	26.85	22.85
+ δ-mem (SSW)	81.70	63.43	**41.41**	37.84	47.05	51.44
+ δ-mem (TSW)	**82.99**	**63.66**	40.40	36.48	46.53	**51.66**
+ δ-mem (MSW)	81.52	60.47	37.37	**38.85**	**49.12**	50.74

1.10× vs. frozen backbone

1.15× vs. best memory baseline

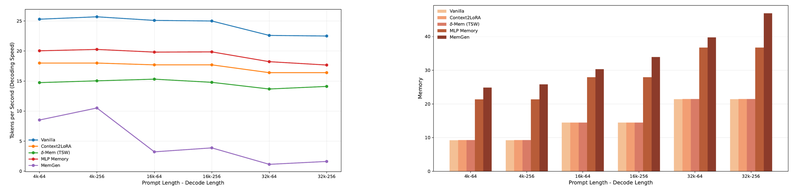
1.31× MemoryAgentBench

1.20× LoCoMo

Consistent across backbones (final average):

Backbone	Frozen	+ C2LoRA	δ-mem SSW	TSW	MSW
Qwen3-4B-Instruct	46.79	44.90	51.44	**51.66**	50.74
Qwen3-8B	47.20	45.92	**50.86**	50.68	50.84
SmolLM3-3B	26.08	34.94	36.67	34.74	**36.96**

Efficiency & Ablations



(a) Decoding throughput. (b) Memory usage.

Figure 3: Inference efficiency under different prompt / decode lengths.

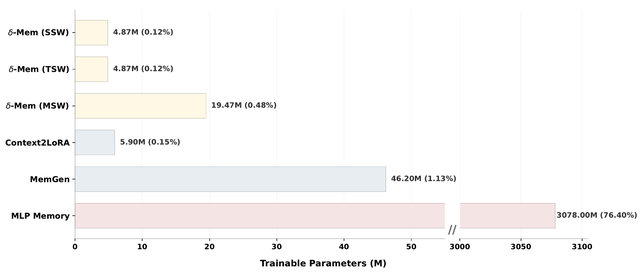


Figure 4: Trainable parameters of memory-augmented methods.

- **4.87M** trainable params (**0.12%**); GPU memory ≈ vanilla even at 32K prompts.
- Ablations: **o** is the best single branch; **qo** ≈ qkvo at lower cost; injecting into **all layers** works best.

A tiny **8×8 state**, written online by the **delta rule**, is enough to steer a frozen LLM!