MemT-1.0: Towards Memory-Native Evolution

Research paper discussion — native state, training, evidence, and tradeoffs

Paper: MemTensor Team Presentation: research discussion

arXiv:2610.26825v2 22 October 2026
https://arxiv.org/abs/2610.26825v2

A bounded state can carry useful history across turns

Central claim

MemT-1.0 writes completed interactions into trajectory-specific matrices; later turns read those matrices instead of receiving the old text.

- Stable token-level writes, snapshot reads, and learned memory projections.
- Mean score: **50.6 Flash** / **47.0 Nano** across 11 memory benchmarks.
- At 128K history, Nano reports **1.86**× end-to-end speedup and **66**× smaller logical persistent state than full-context KV.

Paper, Abstract; Table 2, PDF p. 20; Secs. 6.5–6.6, PDF pp. 22–24.

The design targets the cost of repeatedly carrying history

Full text / retrieval

- Keep detailed evidence available.
- Context, KV state, or stored records grow with history.

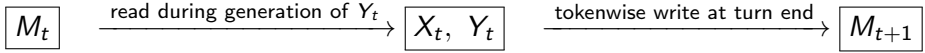
Native memory

- Update a fixed-shape state after each turn.
- Query cost depends on state shape, not stored token count.

Question: How much useful detail survives repeated compression into that fixed state?

Paper, Secs. 1–2 and 8, PDF pp. 5–10, 30.

A turn reads the incoming snapshot, then commits its write



- All queries in turn t use the same incoming M_t.
- The writer processes $X_t \| Y_t$ in token order after the interaction.
- States are reset at trajectory boundaries; backbone and trained memory weights are shared across trajectories.

Consequence

A later token in the same turn does not read a partially updated native state.

Paper, Sec. 2.2 and Fig. 4, PDF pp. 6–8; Sec. 4.2, PDF p. 15.

Memory attention runs beside the backbone attention

$$A_t^{(l)} = A_{\mathrm{ori},t}^{(l)} + A_{\mathrm{mem},t}^{(l)}.$$

- Selected Transformer layers gain a native memory branch.
- **Local memory** $M_t^{(l)}$: dynamic, private to a trajectory.
- **Hyper memory** ϕ: trained projections and gates, fixed online.
- Backbone attention and subsequent backbone computation are preserved.

Evaluated model geometry

Nano: 10 memory layers, 40 MiB FP32 state. Flash: 20 memory layers, 1,280 MiB FP32 state.

Paper, Sec. 2.1 and Fig. 4, PDF pp. 6–7; Appendix B, Table 6, PDF p. 35.

A symmetrized rank-one update writes each token

$$u_i = \sqrt{\beta_i} \odot k_i, \quad z_i = \omega_i \odot v_i, \quad D_i = \mathrm{Diag}(e^{\lambda_i})$$

$$\boxed{S_i = (I - u_i^\top u_i) D_i S_{i-1} + u_i^\top z_i}$$

- k_i is L2-normalized; β_i, ω_i, and λ_i control erase, write, and decay.
- $S_0 = M_t$; the final S_{L_t} becomes M_{t+1}.
- The ordered writes can revise associations made earlier in the turn.

Paper, Sec. 2.2, PDF p. 8; Sec. 4.2, PDF p. 15. Indices for layer/channel omitted.

Symmetry removes an erase-step expansion mode

Original GDN-2 erase

$A = I - k^\top(\beta \odot k)$

Misaligned erase direction and magnitude can increase a vector's norm.

Stable GDN-2 erase

$A_{\mathrm{sym}} = I - u^\top u$

The component along u contracts; orthogonal components do not expand.

With $\|u\|_2^2 < 1$, the erase map is non-expansive. The full recurrence also depends on decay and bounded writes.

Paper, Appendix C.1–C.2, PDF pp. 37–40.

A gated matrix read converts state into token features

$$q_i = \mathrm{L2Norm}(\tilde{h}_i \tilde{W}_Q), \qquad r_i = \frac{q_i M_t}{\sqrt{d_k}}$$

$$A_{\mathrm{mem,i}} = \mathrm{Concat}_h\left[\mathrm{RMSNorm}(r_{i,h}) \odot \mathrm{SoftCapSiLU}(g_{i,h})\right]$$

- Query-dependent readout selects associations from the state matrix.
- The learned output gate sets each channel's influence on generation.
- Soft-cap threshold $\tau = 2$ limits gate magnitude in the evaluated models.

Paper, Sec. 2.2, PDF pp. 8–9; Appendix B.2, PDF p. 36.

MoS expands capacity with dense read and write routing

$$\alpha_{\mathrm{s}} = \mathrm{softmax}_E(\tilde{H} W_{\mathrm{s}} C^\top), \qquad \alpha_{\mathrm{u}} = \mathrm{softmax}_E(\tilde{H} W_{\mathrm{u}} C^\top)$$

$$A_{\mathrm{mem,i}} = \sum_{e=1}^{E} \alpha_{\mathrm{u},i,e} A_{\mathrm{mem},i,e}$$

- Separate routers share trainable expert addresses C.
- Storage weights scale erase and write gates, not log-decay.
- Every expert participates; more state also means more compute.
- Main evaluated Nano/Flash checkpoints use $E = 1$ per layer; $E = 2/4$ belongs to a separate capacity study.

Paper, Sec. 2.3 and Fig. 5, PDF pp. 9–10; Appendix B, Table 6, PDF p. 35.

Training asks memory to imitate a full-history teacher

Teacher: frozen backbone

Sees textual history $C_{<t}$ plus current query and response prefix.

Student: MemT-1.0

Sees current query and prefix; earlier turns are represented by $\mathcal{M}_t$.

$$\ell_{t,k} = T_d^2 \sum_{v \in \mathcal{K}_{t,k}} \hat{q}_{t,k}(v) \log \frac{\hat{q}_{t,k}(v)}{p_{t,k}^{T_d}(v)}, \quad |\mathcal{K}_{t,k}| = 64, \;\; T_d = 1.$$

The teacher is renormalized on Top-64; the student remains normalized over the full vocabulary. Both use the same fixed response prefix.

Paper, Sec. 3.3, PDF pp. 12–13.

CPT builds memory language modeling; post-training teaches answers

Stage	Input to student	Supervision
CPT: 20B tokens	Current 2K-token segment plus incoming state	Next-token cross-entropy; earlier segments only in state
Post-training: 14B tokens	Current query and response prefix plus state	Top-64 teacher KL on response positions

Historical tokens have no direct response loss, yet prediction gradients backpropagate through earlier memory writes. Backbone weights stay frozen during memory training.

Paper, Secs. 3.2–3.6, PDF pp. 11–14; Appendix B.3, PDF p. 37.

Windowed execution makes dense experts trainable

- Chunked WY groups ordered rank-one writes into GPU-friendly work.
- Snapshot reads are batched independently of the writer chunks.
- Complete-step windows bound live MoS projections; backward recomputes intermediates.

Single MoS block, synthetic 32K-token pack, four experts

Peak allocated memory: **47.83** $\rightarrow$ **9.49 GiB** with a 4K-token window.
Median forward + backward: **463.0** $\rightarrow$ **470.6 ms**.

This is a single-block workload, not an end-to-end model-training benchmark.

Paper, Sec. 4.2, Table 1, PDF pp. 15–16; Appendix F, PDF pp. 60–62.

The evaluations separate memory, general ability, and cost

- **Memory:** 11 benchmarks; history written chronologically; queries see state but no prior text. Report an equal-weight mean.
- **General ability:** six benchmarks; compare MemT with its own backbone under matched irrelevant-information conditions.
- **Cost:** BEAM histories; one 512-token query, forced 128-token response; separate history preparation and query TTFT.

Comparison groups

Textual memory: full context, retrieval, memory systems. Parametric memory: Temp-LoRA, δ-mem, and others.

Paper, Sec. 6.1, PDF pp. 18–19; Sec. 6.5, PDF pp. 22–23; Appendix D, PDF pp. 40–53.

MemT leads the evaluated parametric group, but text retains detail

Method	Mean of 11 ↑	LongMemEval-S ↑	BEAM-1M ↑
MemT-1.0-Flash	**50.6**	37.2	37.0
MemT-1.0-Nano	**47.0**	32.6	35.8
Temp-LoRA	38.1	13.4	32.3
MemOS (DeepSeek branch)	68.6	83.2	72.6
Full context (Qwen branch)	66.3	76.8	43.7

Interpretation

Native state works across varied tasks; exact long-history evidence remains harder to preserve.

Paper, Table 2 and Sec. 6.2, PDF pp. 19–20. Systems differ in backbone and mechanism.

General ability is broadly preserved, with task-level regressions

Model vs. own noisy backbone	Six-task mean	IFEval	LiveCodeBench v6
MemT-1.0-Flash	0.0	+0.9	−0.7
MemT-1.0-Nano	−0.9	+3.3	−5.8

- Entries are *score-point differences*, not absolute scores.
- Flash is flat on average; Nano's largest drop is LiveCodeBench v6.
- Irrelevant examples are written into memory for MemT, and prepended to text for the matched backbone condition.

Paper, Table 3, PDF p. 21; Appendix D.4, PDF pp. 51–52.

Token-level writes and post-training matter most in ablation

Nano variant	Mean ↑	Nano variant	Mean ↑
Full model	47.0	Turn-level update	41.0
Without CPT	44.9	Without post-training	36.3
GDN-2, no symmetry fix	46.4	GDN update	46.7

Reading the experiment

Switching to turn-level aggregation costs about 6 points; removing post-training costs about 10.7. Update-rule alternatives are much closer.

Paper, Table 4 and Sec. 6.4, PDF pp. 21–22. Rounded table entries shown.

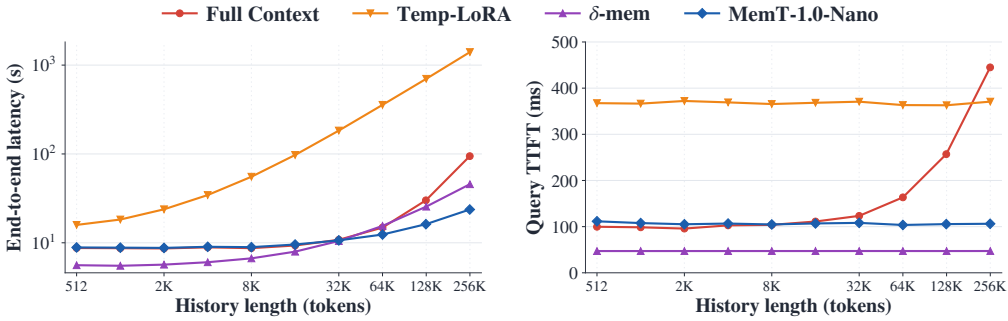
More training and more experts help in controlled studies

- Same Nano post-training trajectory, repeated corpus: mean **36.65** → **45.40** → **46.80** from start to end of epoch 1 to end of epoch 2.
- At matched post-training steps, four MoS experts beat two by **0.41** / **2.01** / **2.69** mean points at 200 / 3,400 / 5,500 steps.

These are fixed-corpus and fixed-backbone comparisons. The capacity study uses different E from the main evaluated checkpoints.

Paper, Sec. 5.3–5.4 and Figs. 6–7, PDF pp. 17–18; Appendix B, Table 6, PDF p. 35.

Nano's long-history advantage includes preparation cost



- End-to-end: about **1.86**× faster at 128K history; short histories favor or match full context.
- Query TTFT after preparation: Nano about 104–112 ms; full context about 257 ms at 128K and 445 ms at 256K.

Paper, Fig. 8 and Sec. 6.5, PDF pp. 22–24. Single request, two GPUs for the HF runs.

Fixed state saves storage, while fidelity has a capacity limit

$$N_{\text{state}} = \sum_{l} E_l H_{kv,l} d_{k,l} d_{v,l}$$

- Nano state: **40 MiB FP32 per trajectory**, independent of accumulated history length at a fixed configuration.
- At 128K history: about **66**$\times$ less logical persistent state than full-context KV cache.
- Shared weights and active-turn KV caches are outside this metric.
- LongMemEval-S and BEAM-1M show the remaining cost in detailed recall.

Paper, Secs. 4.3, 6.2, 6.6, PDF pp. 16, 19–20, 24; Appendix B, Table 6, PDF p. 35.

The agent stores retrieved procedural experience once

1. Summarize successful coding trajectories into **1,040** reusable experiences.
2. Retrieve by issue text with BGE-M3; select experience candidates.
3. Store before execution or when the harness detects a breakpoint.
4. Continue the agent with experience in text or native memory.

Application training

Failed rollouts are resumed with experience; successful continuations form training examples. Main text selects KD with full-suffix supervision.

Paper, Secs. 7.2–7.3 and Fig. 11, PDF pp. 26–28.

Experience improves coding success and repeatability

Benchmark / timing	Experience	Mean success	All 3 succeed
Verified / before	None	66.07%	54.00%
Verified / before	Text	68.87%	61.60%
Verified / before	Native	**69.33%**	**61.60%**
Multilingual / before	None	60.00%	45.33%
Multilingual / before	Text	**64.22%**	**56.33%**
Multilingual / before	Native	62.89%	53.67%

Each task runs three times. Native memory improves on no experience; textual experience remains stronger on Multilingual before execution.

Paper, Table 5, PDF p. 29. Verified: 500 tasks; Multilingual: 300 tasks.

Two boundaries matter when interpreting the agent result

- The experience bank is built and retrieved externally; the tested agent reuses that experience rather than autonomously discovering it.
- Appendix E.2 labels its token selector, 512×512 state, and 105M trainable parameters as **Metis**, whereas the main text describes MemT-1.0-Nano and Appendix B specifies different geometry.
- The application-specific implementation details therefore need author clarification before treating E.2 as a MemT recipe.

Paper, Sec. 7.3 and Sec. 8, PDF pp. 27–30; Appendix B, PDF p. 35; Appendix E.2–E.3, PDF pp. 56–60.

Take-home: a useful bounded memory, with a fidelity frontier

1. **Mechanism:** snapshot reads + stable tokenwise writes make memory a persistent part of model computation.
2. **Evidence:** leading average among tested parametric memory systems, long-history efficiency, and better agent repeatability.
3. **Tradeoff:** fixed state saves per-trajectory storage but loses detailed evidence relative to strong textual memory systems.

Open direction: selective retention, larger effective capacity, and consolidation from local online states into shared trained parameters.

Paper, Secs. 6–9, PDF pp. 18–31.

Backup: evaluated state geometry uses one expert per layer

	Nano	Flash
Backbone	Qwen3.6-35B-A3B	DeepSeek-V4-Flash
Memory layers	10	20
Channels / key / value	16 / 256 / 256	64 / 512 / 512
Experts per layer	1	1
Trainable memory parameters	450.9M	3,608.6M
FP32 state / trajectory	40 MiB	1,280 MiB

Paper, Appendix B, Table 6, PDF p. 35.

Backup: the erase-step stability argument

$$A = I - k^\top(\beta \odot k), \qquad \beta \odot k = \rho k + \gamma e, \quad e \perp k.$$

$$A|_{\{k,e\}} = \begin{bmatrix} 1-\rho & -\gamma \\ 0 & 1 \end{bmatrix}, \qquad \|eA^\top\|_2^2 = 1 + \gamma^2.$$

Symmetrization uses $u = \sqrt{\beta} \odot k$ and $A_{\mathrm{sym}} = I - u^\top u$. Under $\|u\|^2 < 1$, it contracts along u and is unchanged on $u^\perp$.

Paper, Appendix C, PDF pp. 37–40.

Backup: distinguish score, latency, and storage metrics

- Memory benchmark mean: equal-weight mean of 11 benchmark scores; PS-Bench uses $100(1 - \mathrm{ASR})$.
- Query TTFT excludes history preparation; end-to-end latency includes it.
- Storage is logical persistent state after preparation, before query; shared weights and active workspaces are excluded.
- Agent "all 3 succeed" is pass^3; pass@3 is at least one success.

Paper, Sec. 6, Table 2 and Fig. 8, PDF pp. 18–24; Table 5, PDF p. 29; Appendix D, PDF pp. 40–53.