Metis: Memory Foundation Model

Learning persistent memory inside a frozen language model

Paper by Zeyu Zhang et al.

Independent research discussion

arXiv:2607.26760v2 (4 August 2026)
Reading: 9 October 2026

Fixed-size states support memory without replaying history,
with substantial limits in capacity and controllability.

Memory must bridge when facts arrive and matter

Research question

Can a model learn what to store, revise, forget, and use across interactions through its own forward computation?

External textual memory

- Store history as text.
- Retrieve and construct a prompt.
- Optimize retrieval and answering through separate stages.

Paper, Secs. 1–2, PDF pp. 1–6.

Metis

- Persist a compact internal state.
- Read it in the model forward pass.
- Learn memory procedures from future-answer supervision.

Online state changes; learned weights stay frozen

$$(Y_t, \mathbf{M}_{t+1}) = f_\Phi(X_t, \mathbf{M}_t)$$

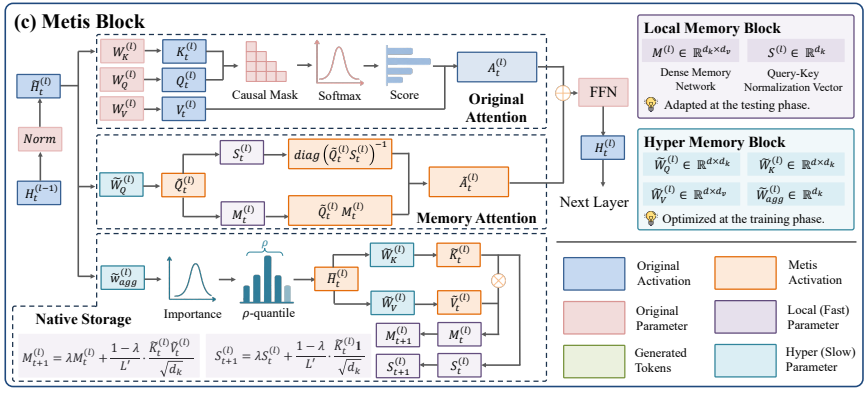
- X_t: current input; Y_t: current response.
- $\mathbf{M}_t$: persistent state, with normalization state $\mathbf{S}_t$ suppressed here.
- Φ: shared backbone and learned memory read/write weights.

Two different kinds of adaptation

Mid-training: freeze the backbone and train memory parameters.
Inference: freeze all learned weights; update state by forward computation.

Paper, Secs. 2.1–2.3 and 5.1, PDF pp. 3–5, 16–17; Appendix A, p. 33.

A layer reads old memory and writes the next state



The original attention branch handles the current step. Memory readout adds history; the write branch prepares future state.

Paper, Fig. 2(c), PDF p. 7, cropped from the author figure. The diagram shows LU; the final model uses GDU.

Select informative tokens before compressing them

$$\mathbf{p}_t = \mathrm{Softmax}\left(\tilde{\mathbf{H}}_t \tilde{\mathbf{w}}_{\mathrm{agg}} / \tau\right)$$

$$L'_t = \mathrm{clip}\left(\min\left\{k : \sum_{r=1}^{k} p_{(r)} \geq \rho\right\}, K_{\min}, L\right)$$

- Sort importance scores; retain enough positions to cover mass ρ.
- Gather L'_t hidden states and project them to $\tilde{\mathbf{K}}_t, \tilde{\mathbf{V}}_t$.
- A straight-through estimator trains the scorer through the dense distribution.

The key distinction

Top-ρ controls probability mass, not a fixed fraction of tokens.

Paper, Sec. 3.3, PDF pp. 8–9. Layer indices omitted.

The linear update exposes the storage trade-off

$$\mathbf{M}_{t+1} = \lambda \mathbf{M}_t + \frac{1-\lambda}{L'_t\sqrt{d_k}}\,\tilde{\mathbf{K}}_t^\top \tilde{\mathbf{V}}_t$$

$$\mathbf{S}_{t+1} = \lambda \mathbf{S}_t + \frac{1-\lambda}{L'_t\sqrt{d_k}}\,\tilde{\mathbf{K}}_t^\top \mathbf{1}$$

- $\mathbf{M}_t \in \mathbb{R}^{d_k \times d_v}$ stores key–value associations; $\mathbf{S}_t \in \mathbb{R}^{d_k}$ tracks normalization.
- Fixed state size; old and new associations share the same matrix.
- Global decay alone does not implement selective semantic forgetting.

Final Metis uses Gated Delta Update (GDU). The paper does not expand its complete recurrence.

Paper, Eqs. (2)–(3), Sec. 3.3, PDF p. 9; Appendix C, pp. 35–36.

A separate query learns how to use historical state

$$\tilde{\mathbf{Q}}_t = \tilde{\mathbf{H}}_t \tilde{\mathbf{W}}_Q, \qquad \tilde{\mathbf{A}}_t = \mathrm{diag}(\tilde{\mathbf{Q}}_t \mathbf{S}_t)^{-1} \tilde{\mathbf{Q}}_t \mathbf{M}_t$$

$$\mathbf{A}_t = \gamma\, \mathbf{A}_{\mathrm{current}} + (1-\gamma)\, \mathrm{Norm}(\tilde{\mathbf{A}}_t)$$

- Learned $\tilde{\mathbf{W}}_Q$ can suppress irrelevant historical keys.
- Query–key normalization controls the readout denominator.
- Output normalization aligns the scale of the two branches.

Implementation adds a unit vector to the denominator for stability; the displayed paper equation omits it.

Paper, Eqs. (4)–(5), PDF p. 9. Current-step attention is retained.

Supervision teaches operations and their side effects

Primary: 357,137 samples

- Remember: fact $\rightarrow$ query.
- Update: old $\rightarrow$ new $\rightarrow$ query.
- Forget: fact $\rightarrow$ revoke $\rightarrow$ query.
- Reflect: facts $\rightarrow$ composed answer.

27 source benchmarks; 406.1M tokens.
Explicit, implicit, and distractor variants.

Auxiliary: 609,443 samples

- Bind similar facts to different entities.
- Forget one fact while retaining another.
- Return to ordinary conversation.
- Ignore memory when irrelevant.

Paper, Sec. 4, Tables 1–4, PDF pp. 13–16. Reflect means composition of stored facts in these tasks.

Three objectives share the same answer likelihood

$$\ell(s,t) = -\frac{1}{|Y_t|}\sum_k \log P(y_{t,k} \mid X_t, Y_{t,<k}; \theta_t)$$

$$\mathcal{L} = \sum_\tau \pi_\tau(e)\, \mathbb{E}_{s\sim\mathcal{D}_\tau}\left[\sum_{t\in\mathcal{Q}_s} \ell(s,t)\right]$$

- **Reconstruction:** recover the earlier passage from state.
- **Operations:** answer according to the revised memory.
- **Regularization:** avoid interference and memory pollution.

Five data subsets; annealed sampling weights $\pi_\tau(e)$. Only selected answer steps are directly supervised.

Paper, Sec. 5, Eqs. (12)–(17), PDF pp. 16–18.

No context means no history replay at query time

Setting	Historical information available to the query
Full Context	All evidence text supplied by the benchmark setting
Partial Context	DenseRAG top-5 sentence chunks
No Context	State or adapted weights after an information phase

- Each instance resets state, ingests evidence, then answers questions.
- MemOps Gold supplies oracle turns; LoCoMo Gold supplies evidence sessions.
- Main scores: per-example median of 3 judge calls, then micro-average.

Paper, Sec. 6.1 and Appendices E–F, PDF pp. 18–19, 37–40. Judge: gpt-4.1-mini; scores allow partial credit.

Memory-only gains leave a large full-context gap

Method	MemOps	Metis Test	LoCoMo	NextMem
Qwen3.5-4B, no history	1.65	16.96	0.18	11.86
Temp-LoRA-4B	8.85	19.34	9.99	24.20
δ-Mem	4.38	15.03	10.79	20.42
Metis-4B	**17.84**	**56.72**	**16.31**	**41.69**
Metis-27B	**24.76**	**73.77**	**26.74**	**50.82**
Qwen3.5-27B, full context	87.90	78.87	65.03	78.80

- Metis is strongest on the four main averages within the no-context group.
- Exact evidence remains much more effective on external memory tasks.

Paper, Tables 5–6, PDF pp. 19–20. MemOps and LoCoMo use Gold; scores on a 0–100 scale. δ-Mem uses a different Qwen3-4B backbone.

Forgetting remains hard outside the synthetic test

Metis-27B: Forget score
Metis Test: **77.50**
MemOps Gold: **10.91**
MemOps Full: **3.86**

A Metis-4B timing failure

1. Store Alice's food preference.
2. Ask the model to forget it.
3. The immediate reply repeats it.
4. A later query receives no answer.

Operation-step responses and subsequent memory behavior are separate targets for controllability.

Paper, Table 5, p. 19; Table 12, p. 35; Fig. 5 and Sec. 6.9, p. 26. Case sequence paraphrased.

Token selection is the largest measured contributor

Metis-4B variant	Overall	Relative change
Full model	33.14	–
No auxiliary data (MS+MP)	26.74	−19.31%
Linear update instead of GDU	32.95	−0.58%
Last token instead of adaptive selection	12.93	−60.98%
Reuse original query	29.09	−12.23%
No query–key normalization	23.72	−28.44%

Accurate input selection and controlled memory readout matter more than the update-rule change in this 4B aggregate.

Paper, Table 7, PDF p. 22. Overall: equal-weight mean of four benchmarks. Changes are relative percentages, not percentage points.

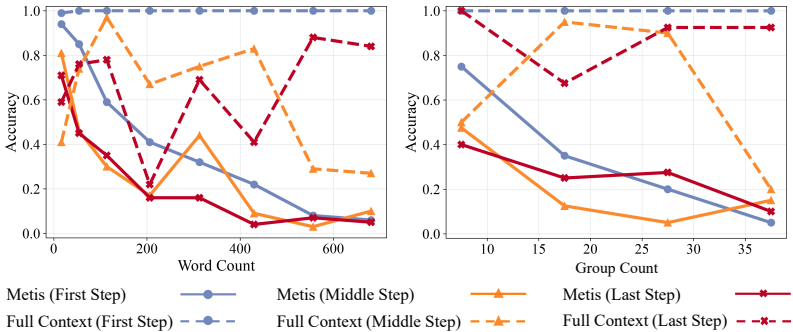
Transfer is strong on ATM and mixed on MemDaily

Method	ATM-Bench Gold	MemDaily Gold
Temp-LoRA-4B / Metis-4B	2.57 / **10.22**	44.92 / **52.54**
Temp-LoRA-9B / Metis-9B	2.86 / **16.49**	**51.42** / 47.29
Temp-LoRA-27B / Metis-27B	2.57 / **18.56**	**59.45** / 59.04

- Neither benchmark was used for training-data construction.
- ATM number questions use exact match; MemDaily uses single-choice accuracy.
- Removing LoCoMo/LongMemEval training sources reduces the six-benchmark mean by 0.76–2.69 points across scales.

Paper, Tables 8–9, PDF pp. 22–23. All methods query without context replay. ATM mixes Jaccard, exact match, and judge scores.

Capacity degrades with both longer writes and history



Left: reset and write all facts once. Right: accumulate state across writes.
Solid: Metis-4B; dashed: full context. Earlier facts weaken; newer facts also fluctuate.

Paper, Fig. 3, p. 23; protocol, pp. 23–24. Left axis counts words. Right Group Count labels do not align with 40 facts grouped by 5; interpret the trend, not an exact update threshold.

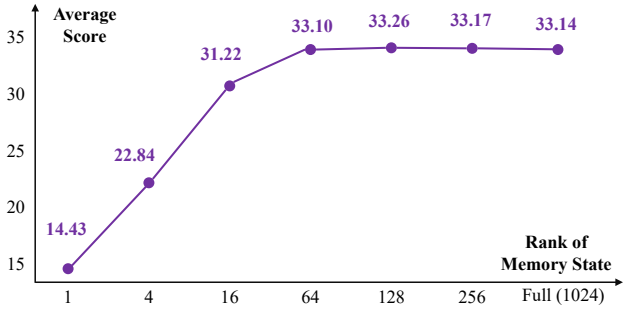
Activating irrelevant memory hurts instruction following

Active stage	Qwen3.5-4B	Metis-4B	Gap (pp)
MMLU-Pro	46.00	40.90	−5.10
IFEval (strict)	76.71	54.53	−22.18
GSM8K	84.53	78.92	−5.61
MMMLU	59.90	56.60	−3.30

- Empty-state scores are close to the original backbone.
- Active stage: store one fixed task-irrelevant evaluation message; the backbone sees the same message in its prompt.
- Frozen weights do not prevent activation-level interference.

Paper, Table 10, p. 24; exact active-stage protocol, Appendix F.4, p. 40.

Rank 64 preserves short-task scores at one-eighth storage



Metis-4B

Full score: 33.14
Rank-64 score: 33.10
Recovery: 99.9%

Full state: 16.79 MB
Rank 64: 2.11 MB

$$\hat{\mathbf{M}} = \mathbf{U}_{:,1:k}\mathbf{\Sigma}_{1:k,1:k}\mathbf{V}^{\top}_{:,1:k}$$

Store the factors, then reconstruct before use. This is a persistence result, not an eightfold compute speedup.

Paper, Fig. 4 / Table 11, p. 25; storage measurements, Appendix G.3, p. 43.

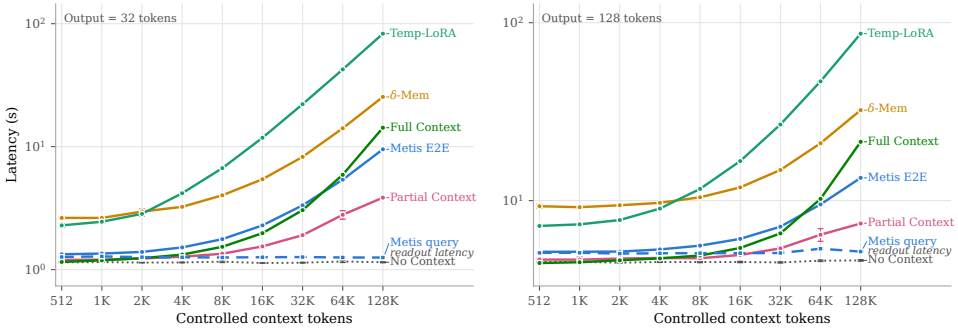
Cold-start latency improves modestly over full context

4B method	Mean E2E (s)	P95 E2E (s)	Mean query (s)
Full Context	0.607	3.012	0.607
RAG top-5	0.268	0.456	0.223
Temp-LoRA	1.567	3.292	0.305
δ-Mem	0.884	1.600	0.656
Metis	0.562	0.926	0.360

- Metis vs. Full Context: mean -7.4%, P95 -69.3%.
- RAG is fastest among history-using methods in this protocol.
- Natural stopping yields unequal output lengths: Metis 8.3 tokens, Full Context 5.7 on average.

Paper, Appendix G.1, Table 16, p. 41. 1,527 LoCoMo Gold instances; state reset and evidence processing included in E2E.

Long-history speedups include a real ingestion cost



First measured E2E crossover: **64K**. At 128K: **1.497**× and **1.595**× faster than Full Context for 32/128 outputs.

Most ingestion time is backbone encoding; the query path stays nearly flat.

Paper, Figs. 7–8, pp. 42–43. Single A800, batch 1, fixed outputs. Generation-only timing: no answer-quality scoring in this sweep.

Users share weights and carry compact separate states

Memory-as-Adapter

- Share frozen model and memory parameters.
- Swap each user's state from storage.
- Stack equal-shape states along the batch dimension.
- Use batched matrix multiplication for readout.

Measured operating points

- At 32K: Full Context 1,118.86 MB/session; Metis 16.79 MB.
- $B = 8$: TTFT 88–91 ms with prebuilt states.
- $B = 64$: 1,585.6 decode tokens/s; 30% latency overhead vs. raw backbone.

Paper, Appendix G.3–G.4, pp. 43–46. Single-node static batches; serving state writes are performed in advance.

A useful memory primitive with three open bottlenecks

1. **Capacity:** fixed state avoids growing caches, but long writes and repeated updates lose information.
2. **Control:** semantic operations emerge, but external forgetting and active-state instruction following remain weak.
3. **Evaluation:** short-task quality and long-history throughput should be tested jointly on persistent, interleaved interactions.

Discussion takeaway

Native memory is a promising learned component for future hybrid memory systems.

Synthesis of Secs. 6 and 8 and Appendix G, PDF pp. 18–28, 40–46. Joint evaluation is a reader-proposed direction.

Paper and discussion sources

Zeyu Zhang et al.
Metis: Memory Foundation Model
arXiv:2607.26760v2, revised 4 August 2026.

- Versioned arXiv record
- Paper PDF
- Project link supplied in the paper
- Model collection supplied in the paper

Source locators refer to 1-based PDF pages. Reused figures are author figures; numerical tables select and reorganize the reported results.

Backup slides: derivation, datasets, supplementary results, serving.

Backup: virtual-prefix attention motivates the state

Exact attention over current tokens and a virtual memory prefix can be split:

$$\mathbf{A} = \mathbf{\Lambda}_{\mathrm{cur}}\mathbf{A}_{\mathrm{cur}} + (\mathbf{I} - \mathbf{\Lambda}_{\mathrm{cur}})\mathbf{A}_{\mathrm{prefix}}.$$

- $\mathbf{\Lambda}_{\mathrm{cur}}$ depends on each query's softmax partition masses.
- Replace the query-dependent matrix by a global mixture weight γ.
- Replace the exponential similarity kernel with a dot product.

$$\mathbf{A}_{\mathrm{prefix}} \;\rightsquigarrow\; \mathrm{diag}\Big(\mathbf{Q}\mathbf{K}^\top\mathbf{1}/\sqrt{d_k}\Big)^{-1}\mathbf{Q}\left(\mathbf{K}^\top\mathbf{V}/\sqrt{d_k}\right)$$

These are architectural approximations. The kernel replacement is not the first-order expansion $1 + z$ of $\exp(z)$.

Paper, Sec. 3.4, PDF pp. 9–11. No uniform softmax-approximation bound is given.

Backup: LU mixes value interference and normalization

Let D_c and $\mathbf{R}_c$ be the target step's denominator and numerator contributions; D_o, R_o collect other steps.

$$\begin{aligned}\tilde{\mathbf{A}} &= (D_c + D_o)^{-1}(\mathbf{R}_c + R_o) \\ &\approx D_c^{-1}\mathbf{R}_c + \underbrace{D_c^{-1}R_o}_{\text{value interference}} \\ &\quad - \underbrace{D_c^{-1}D_oD_c^{-1}\mathbf{R}_c}_{\text{target dilution}} - \underbrace{D_c^{-1}D_oD_c^{-1}R_o}_{\text{cross term}}.\end{aligned}$$

- Expansion requires invertible D_c and small $D_c^{-1}D_o$.
- Independent memory queries can reduce off-target similarities.
- Many individually small off-target terms can still accumulate.

Paper, Sec. 3.5, PDF pp. 11–13. D_c, D_o, R_o are teaching shorthand. This analysis uses LU, not the final GDU recurrence.

Backup: data scale and training progress differ

Primary operation	Samples	Tokens (M)
Remember	56,855	362.0
Forget	136,271	21.7
Update	81,501	11.0
Reflect	82,510	11.4

- Remember contributes about 89.1% of primary tokens.
- 4B: 14K steps / 1 epoch; 27B: 14K / about 0.4 epoch.
- 9B: validation-based early stopping at 8K / about 0.5728 epoch.
- AdamW, learning rate $2 \cdot 10^{-4}$, BF16, 8×H100; backbone frozen.

Paper, Table 2, p. 15; training configuration, Sec. 6.1, p. 19. A further 609,443 auxiliary samples support interference control.

Backup: MemOps Full is harder for native memory

No-context method	4B	9B	27B
Temp-LoRA	9.98	12.19	10.83
Metis	9.70	11.53	15.35
Full-context Qwen3.5 reference	83.52	85.69	87.19

- Full evidence: three complete segments, 24 utterances.
- Metis-4B and 9B do not beat matching Temp-LoRA averages here.
- Metis-27B leads the no-context average, with strong reflection but weak forgetting.

Paper, Appendix B, Table 12, PDF pp. 34–35. Micro-averaged judge scores, 0–100 scale.

Backup: GDU and LU favor different tasks

Scale	Update	LoCoMo Gold	Metis Test	Overall
4B	GDU / LU	16.31 / 11.97	56.72 / 58.54	33.14 / 32.95
9B	GDU / LU	16.81 / 15.37	57.92 / 68.66	34.44 / 35.59
27B	GDU / LU	26.74 / 14.16	73.77 / 75.32	44.02 / 41.50

- GDU leads on LoCoMo at every scale; LU leads on the synthetic test.
- 9B comparison also changes training duration: GDU 8K, LU 14K steps.
- Under LU, removing MS slightly improves Overall (32.95 to 33.67); removing all auxiliary data lowers it to 28.85.

Paper, Appendix C, Tables 13–14, PDF pp. 35–36. Overall is the equal-weight mean of four benchmarks.

Backup: removing sources redistributes performance

Scale	Six-bench mean	Change	LoCoMo change	NextMem change
4B	30.90	−1.66	−3.13	−4.39
9B	32.83	−0.76	−1.09	−1.85
27B	39.59	−2.69	−5.59	+2.49

- Exclude all instances synthesized from LoCoMo and LongMemEval.
- About 2.61% of instances are removed, concentrated in selected task slices.
- Overall capability largely survives; individual tasks move differently.
- Entity coverage and supervision mixture both change.

Paper, Sec. 6.5, Table 9, PDF pp. 22–23. Changes are score points; the six-benchmark mean includes the two OOD benchmarks.

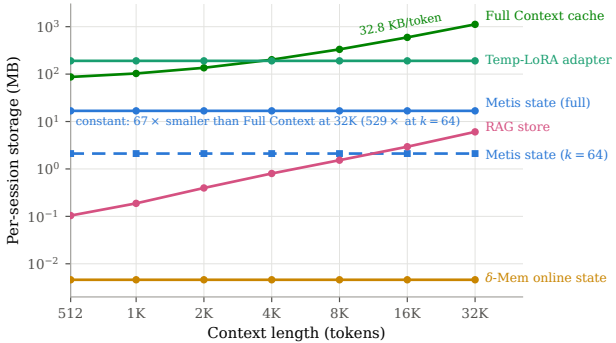
Backup: the architecture transfers across backbones

Backbone	No history	Metis	Full Context
Llama3.1-8B	10.69	36.03	70.61
Gemma4-12B	8.36	37.68	77.30
Gemma4-31B	5.40	33.44	78.18
Llama3.1-70B	12.71	39.29	75.06

- Each backbone receives memory-specific mid-training; checkpoint at 14K steps.
- Every tested family benefits from the internal memory state.
- Scaling is non-monotonic under the shared training recipe.

Paper, Appendix D, Table 15, PDF pp. 36–37. Equal-weight average over the four main benchmarks.

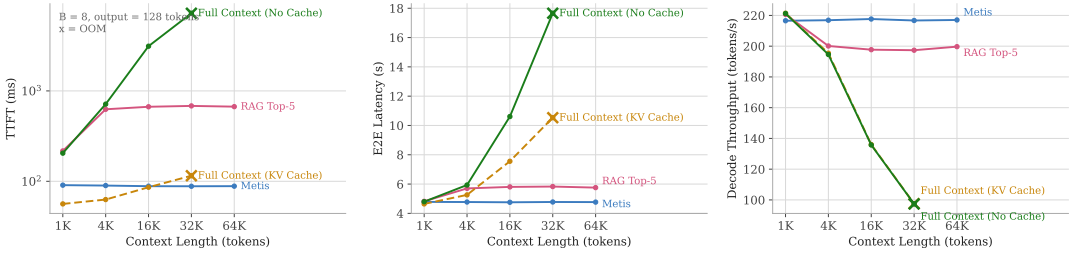
Backup: persistence cost is fixed, but quality also differs



At 32K: Full Context / Metis Full / Metis rank 64 = 1,118.86 / 16.79 / 2.11 MB. δ-Mem is much smaller at 4.6 KB.

Paper, Fig. 9 and Appendix G.3, PDF pp. 43–44. Runtime per-session storage, not total model or peak inference memory. The paper's 128K storage ratios are extrapolations.

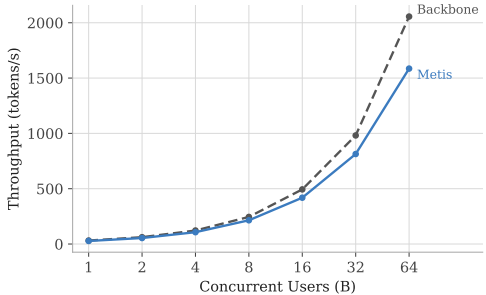
Backup: prebuilt states make the serving path flat

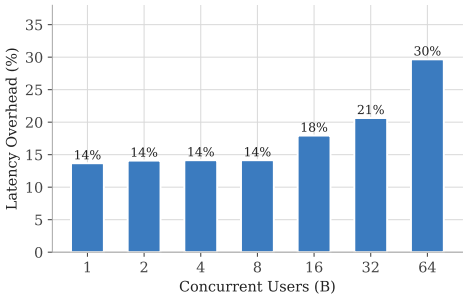


- Batch 8; 128 output tokens/user; 1K–64K historical tokens/user.
- Metis includes loading and state injection, but history writes are prebuilt.
- Full Context includes both resident-KV and no-cache variants; both run out of memory at 64K in this batch.

Paper, Fig. 10 and Appendix G.4, PDF pp. 44–46. Synthetic, single-node, static-batch workload; medians over five runs.

Backup: memory adds overhead while preserving batching





Relative to the raw backbone with identical prompts: about 14% latency overhead through $B = 8$, 30% at $B = 64$.

Continuous batching and interleaved read/write consistency remain future systems work.

Paper, Fig. 11 and Appendix G.4, PDF pp. 45–46. Throughput is aggregate decode throughput; state loading/re-stacking is counted.