Looped Latent Attention

Cross-Loop KV Compression for Looped Transformers

Paper: James O'Neill and Fergal Reid, Fin AI Research

Independent research reading

arXiv:2607.15456v2 (24 July 2026)
Reading: 20 September 2026

Compress the recurrent KV trajectory while preserving loop-specific states.

Weight sharing leaves a request-sized memory bill

The concrete mismatch

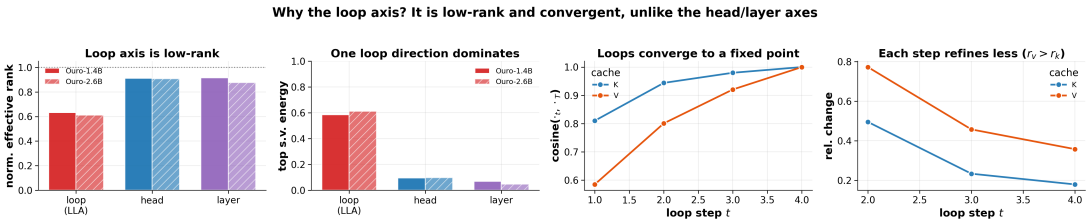
A shared D-layer stack runs T times, but each loop produces its own K/V cache.

$$\text{KV scalars per token} = 2TDHd_{\text{head}}$$

- Ouro-1.4B: $D = 24$, $T = 4$, $H = 16$, $d_{\text{head}} = 128$.
- Parameters of a shared stack; KV state across 96 effective layers.
- Long context and large batches amplify this per-request cost.

Paper, Secs. 1 and 3, PDF pp. 1–4; Appendix A, p. 14.

The strongest redundancy lies along recurrence



- Loop effective rank: roughly 0.61–0.63 of its maximum.
- The first direction carries about 60% of energy; other axes are flatter.
- V stabilizes more slowly than K: allocate $r_v > r_k$.

Original Fig. 2, Sec. 4, PDF p. 5. Convergence alone does not imply low rank.

Store one code; reconstruct the requested loop

For one token, layer, head, and K or V

$$x = [x_1; \ldots; x_T] \in \mathbb{R}^{Td_{\text{head}}}, \qquad c = (x - \mu)W_{\text{down}} \in \mathbb{R}^r$$

$$\hat{x}_t = cW_{\text{up},t}^{\top} + \mu_t$$

- W_{down}: $Td_{\text{head}} \times r$; $W_{\text{up},t}$: $d_{\text{head}} \times r$.
- Separate K and V codes; loop-specific means and reconstruction maps.
- Fit pre-RoPE content; apply the original RoPE after reconstruction.

Paper, Eqs. (1)–(3), Sec. 3, PDF pp. 3–4. Row-vector convention.

The exact saving concerns persistent KV scalars

$$\rho = \frac{2THd_{\text{head}}}{(r_k + r_v)H} = \frac{2Td_{\text{head}}}{r_k + r_v}$$

Worked example: Ouro-1.4B

$T = 4$, $d_{\text{head}} = 128$, $(r_k, r_v) = (96, 160)$:

$$\rho = \frac{1024}{256} = 4, \qquad 768 \text{ KiB/token} \longrightarrow 192 \text{ KiB/token}.$$

- Same precision; fixed codec weights and scratch buffers are excluded.
- Every token retains its own code: cache size still grows with context.

Eq. (4), PDF p. 4; Table 5, p. 11. Numerical substitution is a teaching example.

Teacher geometry initializes a small conversion problem

1. Collect pre-RoPE K/V on 255 calibration prompts and center them.
2. Use top-r right singular vectors of stacked-loop activations.
3. Freeze the teacher; refine only codec parameters.

$$\mathcal{L} = \mathrm{KL}(p_{\text{teacher}} \| p_{\text{swap}}) + \frac{0.5}{TD} \sum_{t,d} \| a_{t,d} - \hat{a}_{t,d} \|_2^2$$

- Typical conversion: 3000 steps, batch 8, length 512, AdamW.
- Select the best held-out KL checkpoint; ablations use 300 steps.

Eq. (5), PDF p. 4; Appendix A, pp. 14–15. $a_{t,d}$ is the attention output.

Compare axes at the same persistent-cache budget

Method	Compressed axis	Role
LLA	Loop	Per-head trajectory code
LLA-2D	Loop and head	One code per layer/token
mla_head	Head	Separate code for each loop
cla	Layer	Adjacent-layer reuse
kv_quant	Precision	Inference-only quantization
Final-loop reuse	Loop endpoint	Zero-parameter control

- LLA-2D saves another factor H only at *fixed ranks*.
- Matched-budget comparisons need different ranks across variants.
- A failed endpoint control is not a direct evaluation of trained MELT.

Secs. 2–3, PDF pp. 2–4; Appendix A, p. 14.

Evaluation separates fidelity, generation, and capacity

- **Teacher-forced fidelity:** held-out KL and top-1 agreement.
- **Free generation:** GSM8K, MATH-500, code, and BBH.
- **Likelihood tasks:** MC suites on given tokens.
- **Systems:** fabricated decode caches pushed to OOM on one H200.

Protocol matters

1.4B iso-cache codecs: 100 GSM8K items; teacher: 1319.
Cold-start sweep: 200 codec items, earlier streaming decoder.
Most quality runs use two-pass reconstruction, storing full K/V.

Sec. 5, PDF p. 6; Appendix A, pp. 14–15; Sec. 7, p. 11.

At 4×, a trajectory code beats endpoint reuse

Ouro-1.4B	KL ↓	GSM8K ↑	MC-avg ↑
Teacher (1×)	—	.794	.646
Final-loop reuse	—	.000	.443
LLA	**.059**	**.800**	.640
Head codec	.122	.522	.585
Layer sharing	.267	.660	.597
KV quantization	—	.702	.645

- Loop states are compressible but functionally distinct.
- Quantization is competitive at 4×; LLA does not win every metric.

Selected rows of Table 1, PDF p. 7. Teacher and codec sample counts differ.

Broad-suite numbers limit the “near-lossless” claim

Ouro-1.4B	GSM8K	HumanEval	MATH-500	MMLU-Pro
Teacher	.794	.707	.74	.481
LLA 1.33×	.795	.756	.70	.467
LLA 2×	.750	.744	.63	.469
LLA 4×	.590	.610	.52	.416

- Light compression is often close; sensitivity remains task-dependent.
- Table 7 contradicts the Fig. 5 caption’s “code/MMLU-Pro through 4×” claim.
- The .590 and .800 GSM8K results belong to different protocols.

Table 7, PDF p. 16; Fig. 5/Table 1, p. 7; Appendix A, p. 14. HumanEval: base pass@1.

The loop-axis advantage also appears at 2.6B

GSM8K EM	4×	5.33×	10.67×
LLA	.834	.817	.821
Head codec	.585	.772	.775
KV quantization	.814	.351	—

- Teacher ≈ .83; LLA and head arms use all 1319 items.
- MC-avg: teacher .653; LLA .645/.646/.640 across these budgets.
- This supports transfer of the ordering, not monotonic gains with compression.

Table 2, PDF p. 8; Table 11, p. 19; Appendix A, p. 14.

SVD initialization drives the largest ablation gain

4× codec, 300-step conversion	Held-out KL ↓
SVD + ranks 96/160 + attention matching	.105
First-loop SVD initialization	.128
Random initialization	.771
Equal ranks 128/128	.129
KL only	.110

- Random initialization has about 7.3× the reference KL.
- More V capacity helps; attention-output matching adds a smaller gain.
- No multi-seed significance estimate is supplied for these differences.

Table 10, Fig. 12, PDF p. 18.

Huginn transfers the idea to 32 recurrent steps

Compression	SVD KL	SVD top-1	Distilled KL	MC-avg
16×	.005	.970	—	.523
32×	.010	.954	—	.525
64×	.197	.799	.145	.486
128×	.763	.620	.177	.473

- Teacher MC-avg: .523; SVD alone is strong through 32×.
- Distillation recovers fidelity at higher compression.
- These are decoder-independent results, not long-generation or serving tests.

Table 3, Sec. 5, PDF p. 9. Top-1 means token agreement with the teacher.

On-policy distillation visits the deployed model's prefixes

$$\mathcal{L}_{\mathrm{off}} = \mathbb{E}_{y \sim p_T}\left[\frac{1}{|y|}\sum_t \mathrm{KL}\big(p_T(\cdot \mid y_{<t}) \| p_S^\theta(\cdot \mid y_{<t})\big)\right]$$

$$\mathcal{L}_{\mathrm{on}} = \mathbb{E}_{\tilde{y} \sim \mathrm{sg}(p_S^\theta)}\left[\frac{1}{|\tilde{y}|}\sum_t \mathrm{KL}\big(p_T(\cdot \mid \tilde{y}_{<t}) \| p_S^\theta(\cdot \mid \tilde{y}_{<t})\big)\right]$$

- Same frozen teacher target; change the prefix distribution.
- Stop gradients through sampling; update only the codec.
- No reward or correctness filtering is used.

Eqs. (6)–(7), Sec. 6, PDF pp. 9–10. T: teacher; S: student.

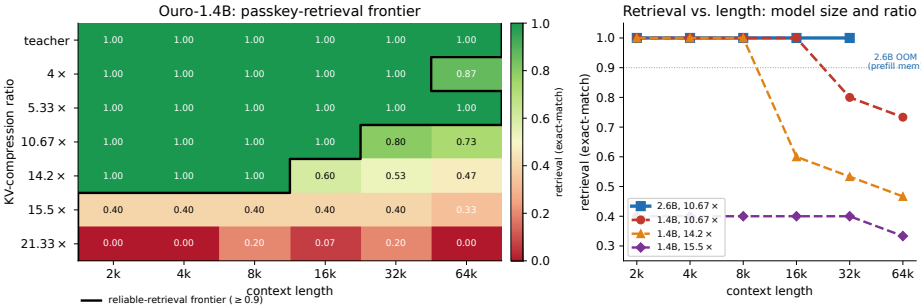
Fewer missing answers need not mean higher accuracy

MATH-500	Off-policy	On-policy	No-answer: off → on
3.2×	.46	.70	.21 → .06
4×	.43	.66	.15 → .03
5.33×	.53	.69	.12 → .03
10.67×	.65	.69	.09 → .07
21.33×	.66	.62	.11 → .05

- At 4×: +23 percentage points in final-answer accuracy.
- At 21.33×: fewer missing answers, but lower accuracy.
- Rollout correction cannot guarantee recovery of lost information.

Table 4, PDF p. 10. Teacher accuracy .74; main codec evaluations use 200 items.

Compression reduces the usable retrieval horizon



- 10.67×: perfect to 16k; .80 at 32k and .73 at 64k.
- Exception to the prose: 4× at 64k is .87, below the .9 threshold.
- 15 trials per cell; 2.6B's 64k OOM is a memory failure.

Original Fig. 8, PDF p. 11; Table 12, p. 20. Latent-store decoder.

Capacity improves while current throughput falls

Ouro-1.4B, one H200	Teacher	LLA 4×	LLA 21.3×
Persistent KV / token	768 KiB	192 KiB	36 KiB
Max batch at 4096 tokens	32	128	768
Full-KV fast-path ms/token	63.6	74.1	—
Capacity-path peak tokens/s	748	—	108

- Full-KV fast path keeps reconstructed states: no memory saving.
- Latent-store path saves memory but reconstructs on every read.
- Batch capacity is measured with fabricated decode caches to OOM.

Table 5, Sec. 7, PDF p. 11. These rows describe distinct execution paths.

Absorbed attention exposes a speed–fidelity gap

$$q^\top (W_{\mathrm{up}} c) = (W_{\mathrm{up}}^\top q)^\top c$$

- Moving projections outside attention can avoid explicit reconstruction.
- Full RoPE blocks direct absorption; a decoupled branch changes fidelity.

Kernel geometry benchmark

Random weights; 262k context.
Absorbed: 4.25 ms/token.
Teacher: 4.46 ms/token.
2.30× faster than reconstruction.

Post-hoc fidelity: KL

Full RoPE: .059.
Decoupled: .55.
With query adapter: .34.
Extended training: .30.

Appendix J, Tables 13–14, PDF pp. 20–21. Column-vector identity; speed is a microbenchmark.

Three evidence boundaries govern the interpretation

1. **Preserve protocols.** Checkpoints, decoders, and evaluation subsets differ across tables. Broad-suite GSM8K repeats cold-start numbers despite different protocol descriptions.
2. **Check figures against tables.** The 4× code/MMLU-Pro claim and the blanket 64k retrieval claim are too strong.
3. **Separate system goals.** Exact KV storage savings, sampled-task quality, and random-weight kernel speed are distinct results.

Reading assessment based on Secs. 5–7 and Appendices A, C, I, J. No independent rerun.

Loop trajectories are a useful compression target

What the paper establishes

A learned low-rank cross-loop codec can preserve useful behavior while substantially reducing persistent KV state in the tested recurrent models.

- SVD exploits teacher geometry; loop-specific reconstruction matters.
- On-policy conversion helps long rollouts at several budgets.
- Capacity gains are real; faithful end-to-end speed remains unresolved.

James O'Neill and Fergal Reid. *Looped Latent Attention: Cross-Loop KV Compression for Looped Transformers*.
arXiv:2607.15456v2, 24 July 2026.

Read: full PDF, main_minimal.tex, results_space_optimized.tex, figures, and references.

Backup: cold-start ranks define a separate frontier

Compression	(r_k, r_v)	GSM8K	Held-out train-KL
4.00×	(96,160)	.590	.0160
3.20×	(120,200)	.575	.0153
2.67×	(144,240)	.575	.0157
2.00×	(192,320)	.750	.0155
1.60×	(240,400)	.755	.0070
1.33×	(288,480)	.795	.0035

Teacher reference .820 on 100 items; codecs on 200 items.
Earlier streaming reconstruction decoder; do not merge with Table 1.

Table 8, PDF p. 17; Appendix A, p. 14.

Backup: the full memory path needs two passes

1. Capture teacher pre-RoPE K/V across loops.
2. Encode and reconstruct the captured vectors.
3. Re-forward with reconstructed K/V injected before RoPE.

What is retained?

Quality path: full post-RoPE reconstructed K/V.
Capacity path: latents, with reconstruction on read.
Long prefill: stream layer-wise reconstruction and free captures.

The mathematical codec does not establish a cost-free single-pass decoder.

Appendix A, PDF p. 14; Appendix J, p. 20.

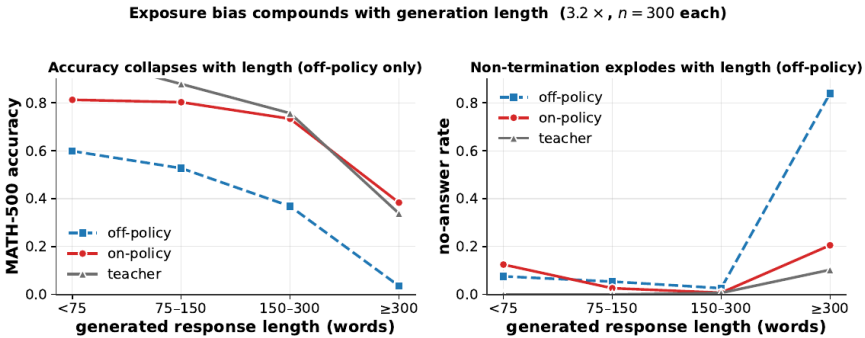
Backup: token eviction composes in storage

260 retained tokens: sink 4 + window 256	GSM8K
Eviction on a near-lossless 8-bit codec	.510
LLA 4× + eviction	.520

- About 74% of the few-shot prompt is evicted; quality cost is substantial.
- At 21.3×, 1024 kept tokens give $36\,\mathrm{KiB} \times 1024 = 36\,\mathrm{MiB}$.
- A 1M-token input with eviction is not lossless storage of 1M tokens.

Table 9, PDF p. 17; Appendix J, p. 20.

Backup: length bins reveal a rollout failure mode



On-policy tracks the teacher more closely at 3.2×. Generated length also reflects task difficulty; this plot alone is not a controlled causal test.

Original Fig. 7, PDF p. 10; figure labels $n = 300$ per arm.