StateLoop: Persistent State

for Looped Transformers

Paper by Feichen Song et al.
Research group paper discussion

arXiv:2610.35544v1 28 October 2026

Publication not found as of 2026-10-28
Public code not found as of 2026-10-28

The problem: one hidden stream must compute and remember

- A looped Transformer reuses one backbone T across K iterations.
- Vanilla feedback sends each backbone output directly into the next iteration.
- Useful early information may need to survive several later transformations.

StateLoop's design

Keep the direct backbone stream and add one gated persistent state per token, carried across the *loop-depth* axis.

Paper, Secs. 1–2, PDF pp. 1–2. Persistence here is across internal loop iterations.

Each token has a direct path and a state path

Direct backbone stream

$$H^k \longrightarrow Z^{k+1} \xrightarrow{T} H^{k+1}$$

The current backbone output remains available to the next iteration.

Persistent state stream

$$S^k \xrightarrow[H^k]{\text{gated update}} S^{k+1}$$

Each token selectively carries state across loop iterations.

The streams meet before the next loop

$$Z^{k+1} = H^k + \text{Read}(S^{k+1})$$

Paper, Fig. 1 and Sec. 3.1, PDF p. 3. Backbone and state-update parameters are shared across iterations.

The gate chooses how much old state to retain

$$U^k = \mathrm{RMSNorm}_h(H^k), \qquad P^k = U^k W_u^\top$$

$$\Delta^k = \mathrm{softplus}(U^k W_\Delta^\top + b_\Delta), \qquad g^k = \exp[-\exp(A_{\log}) \odot \Delta^k]$$

$$\boxed{\widetilde{S}^{k+1} = g^k \odot S^k + (1 - g^k) \odot P^k}$$

- $g^k \approx 1$: retain; $g^k \approx 0$: write from this backbone output.
- Each token and state channel can have a different gate.

Paper, Sec. 3.2.1, PDF p. 4, Eqs. (5)–(6).

The state is read without replacing the backbone stream

$$S^{k+1} = \mathcal{X}_\psi(\widetilde{S}^{k+1}), \qquad N^{k+1} = \mathrm{RMSNorm}_s(S^{k+1})$$

$$\boxed{Z^{k+1} = H^k + N^{k+1} W_{s\to z}^\top}, \qquad H^{k+1} = T(Z^{k+1})$$

$$Y = H^{K-1} + N^K W_{s\to out}^\top$$

- One projection feeds the next loop; a separate projection feeds final prediction.
- Standard next-token cross-entropy trains the full model end-to-end.

Paper, Secs. 3.1 and 3.3, PDF pp. 3–5, Eqs. (3)–(4), (10)–(12).

State update and token mixing act on different axes

Across loop depth

Every variant uses the same selective gate to update each token's state.

Variant	Mixing across tokens
StateLoop	Identity
+ Attn	Causal attention
+ Mamba-2	Residual Mamba-2 block

Autoregressive decoding

The identity variant keeps a transient state for the current token across its K iterations. The token-mixing variants also need loop-specific caches.

Paper, Secs. 3.2.2 and 4.4, PDF pp. 4 and 7–8.

The comparison fixes data and backbone at 1.4B

- Pile pretraining from scratch: 30B tokens, 2,048-token sequences, batch 256.
- Shared 24-layer LLaMA-style backbone; width $d = 2048$, $K = 4$.
- StateLoop state width $d_s = 2048$; three mixer variants.
- Controls: Vanilla Looped, wider FFN, and alternative loop connections.

Metrics

Pile and LAMBADA token PPL; WikiText-2 word PPL; eight-task zero-shot and five-shot accuracy; decoding latency in ms/token.

Paper, Sec. 4.1, PDF p. 5. MeSH and Parcae match 96 layer applications but retain their own structures.

Persistent state lowers language-modeling perplexity

1.4B model	Params (M)	Pile	LAMBADA	WikiText-2
Vanilla Looped	1420.92	7.79	6.94	16.50
Vanilla + FFN	1467.22	7.75	7.00	16.34
MeSH	1421.51	7.74	6.77	16.43
StateLoop	1437.71	7.61	6.71	15.96
StateLoop + Attn	1454.49	**7.60**	**6.64**	**15.93**
StateLoop + Mamba-2	1467.23	**7.60**	6.88	15.95

- Identity StateLoop adds 1.18% parameters and reduces Pile PPL by 0.18.
- A parameter-matched wider FFN does not reproduce the gain.

Paper, Table 1, PDF p. 6. Lower PPL is better; WikiText-2 is word PPL.

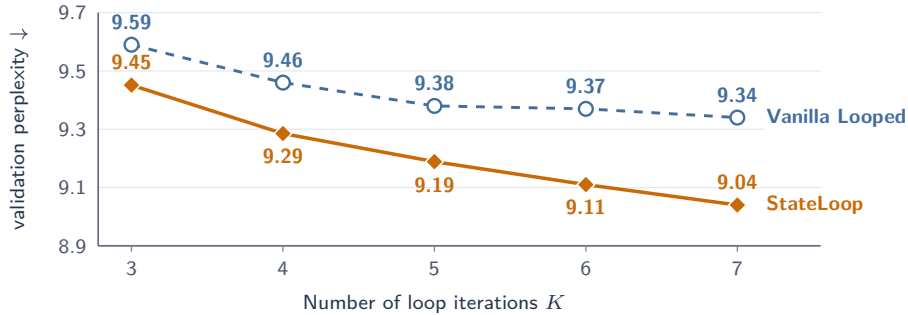
Eight-task average accuracy rises under both prompt settings

1.4B model	Zero-shot Avg-8 (%)	Five-shot Avg-8 (%)
Vanilla Looped	53.31	55.12
Vanilla + FFN	52.85	55.54
StateLoop	54.28	56.60
StateLoop + Attn	54.74	**57.15**
StateLoop + Mamba-2	**54.75**	57.13

- Identity StateLoop: +0.97 zero-shot and +1.48 five-shot percentage points.
- Average gains do not mean every task improves; zero-shot ARC-Easy dips slightly.

Paper, Table 2, PDF p. 7. Avg-8 is the unweighted mean of eight task accuracies.

The StateLoop advantage grows with trained loop depth



- Pile PPL gap: 0.14 at $K = 3$, 0.30 at $K = 7$.
- Every plotted K is a separately trained 410M, 25B-token pair.

Paper, Sec. 4.3 and Fig. 2, PDF pp. 6–7. Both methods improve with more loops.

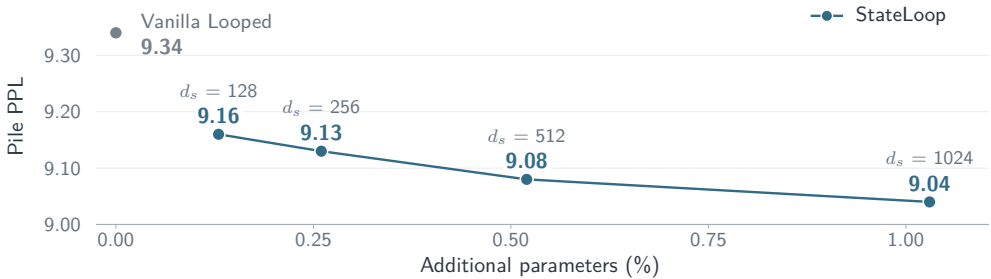
The state branch costs less than one more backbone loop

1.4B decoding	K	ms/token	Change vs. Vanilla $K = 4$
Vanilla Looped	4	121.224	baseline
StateLoop	4	122.943	+1.418%
StateLoop + Attn	4	123.778	+2.106%
StateLoop + Mamba-2	4	123.397	+1.792%
Vanilla Looped	5	151.527	+24.997%

- The comparison measures state-branch cost against a fifth full iteration.
- BF16, batch 1, 1,024-token prefill + 1,024-token continuation.

Paper, Table 3, PDF p. 8. Means over five runs on one GPU.

Even a narrow state preserves much of the gain



- Vanilla: PPL 9.34; $d_s = 128$: 9.16; $d_s = 1024$: 9.04.
- Halving state width from 1,024 to 512 yields PPL 9.08.

Paper, Sec. 5.1 and Fig. 3, PDF p. 8. 410M, $K = 7$, 25B tokens, identity mixer.

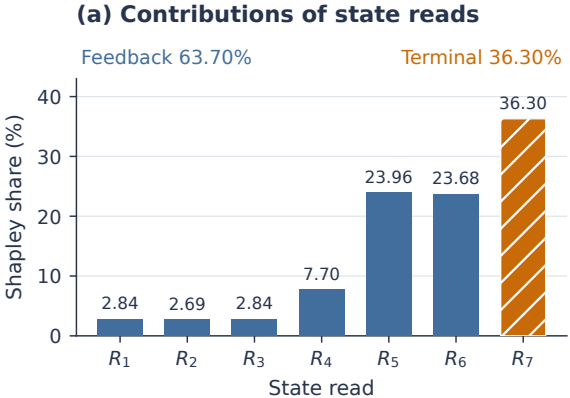
Ablations separate the branch, persistence, and update rule

Intervention	Variant	Params (M)	Pile PPL
Feedback	State-only next input	409.00	9.44
	StateLoop direct path	411.10	**9.04**
Persistence	Reset state each loop	411.10	9.25
	Persist across loops	411.10	**9.04**
Update	Vanilla Looped	406.90	9.34
	StateLoop with GRU	416.34	9.09
	StateLoop selective SSM	411.10	**9.04**

- The branch helps without persistence; persistence adds a further 0.21 PPL gain.
- GRU retains most of the gain: the result is not unique to one gate formula.

Paper, Sec. 5.2 and Table 4, PDF p. 9. 410M, $K = 7$, 25B tokens.

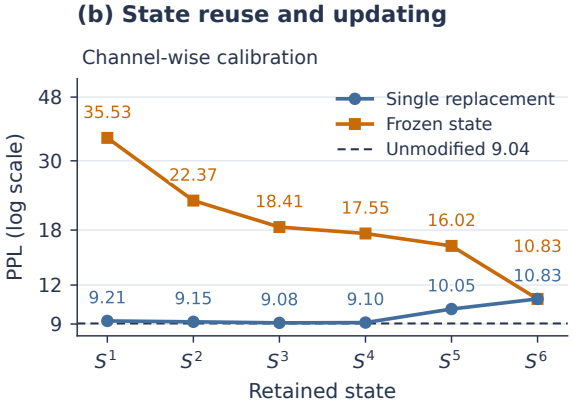
Late state reads receive most attribution



- Mask all 2^7 combinations of state reads.
- Feedback reads: 63.70% of attributed NLL increase.
- Terminal read: 36.30%.
- R_5, R_6, R_7 dominate.

Paper, Sec. 5.3 and Fig. 4(a), PDF p. 10. Shapley shares refer to the read-masking intervention.

Earlier states can be reused once; freezing hurts



- Calibrate an earlier state's channel mean and scale.
- Replace S^4 once using S^3: PPL 9.08.
- Freeze reuse from S^3: PPL 18.41.
- Unmodified model: 9.04.

Paper, Sec. 5.3 and Fig. 4(b), PDF p. 10. Backbone computation continues in both interventions.

The evidence supports a specific design claim

1. A direct backbone path plus a gated extra state beats state-only feedback.
2. Resetting the state loses part of the gain at equal parameter count.
3. Later reads and continued updates contribute materially to prediction.

Scope of the reported evidence

Main 1.4B comparisons use $K = 4$; systematic depth and state interventions use 410M models. Decoding cost is measured under a specified batch-1 protocol.

Paper, Tables 1–4 and Figs. 2–4, PDF pp. 6–10; Sec. 6, p. 11.

Take-home: explicit state helps recurrent depth

- StateLoop retains a small, gated state per token across loop iterations.
- It improves Pile PPL and eight-task average accuracy with modest added latency.
- Its PPL advantage grows with trained depth in the 410M study.
- The next test is the quality–latency tradeoff at larger depths and scales.

Paper: Feichen Song, Boyi Zeng, He Li, Shixiang Song, Yiqin Hao, Zitong Wang, Yizhou Yang, Yifan Liu, Ziwei He, Zhouhan Lin. *StateLoop: Persistent State for Looped Transformers*. arXiv:2610.35544v1.

Publication not found as of 2026-10-28; public code not found as of 2026-10-28.

Backup: initialization and next-token training

$$E = \sqrt{d}\,\mathrm{Embed}(x), \qquad S^0 = 0, \qquad H^0 = T(E)$$

$$p_\theta(\cdot \mid x_{\le t}) = \mathrm{softmax}(\mathrm{LMHead}(Y_t))$$

$$\mathcal{L}_{\mathrm{LM}} = -\frac{1}{n-1}\sum_{t=1}^{n-1} \log p_\theta(x_{t+1} \mid x_{\le t})$$

- $H^k \in \mathbb{R}^{n\times d}$; $S^k \in \mathbb{R}^{n\times d_s}$.
- The whole network is trained end-to-end with ordinary causal LM loss.

Paper, Secs. 3.1 and 3.3, PDF pp. 3–5, Eqs. (1), (11)–(12).

Backup: benchmark interpretation

- Eight-task Avg-8 is an unweighted mean of accuracy percentages.
- Pile and LAMBADA PPL are token-level; WikiText-2 PPL is word-level.
- MeSH and Parcae match the 96 backbone layer applications but keep their original partial-loop designs.
- StateLoop + Attn and + Mamba-2 add token-axis mixing and loop-specific caches.

Paper, Secs. 4.1 and 4.4, PDF pp. 5 and 7–8.