Understanding and Improving Length Generalization in Recurrent Models

Paper discussion: training on attainable states

Paper authors: Ricardo Buitrago Ruiz and Albert Gu

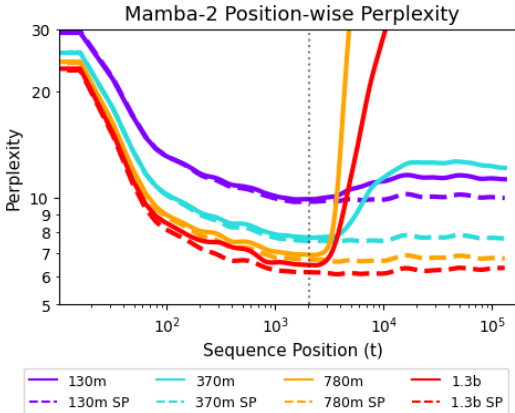
Carnegie Mellon University / Cartesia AI

arXiv:2507.02782v2 | Reading: 12 October 2026

A short training chunk can expose a model to a long-history state.

arxiv.org/abs/2507.02782v2 Independent reading presentation

A recurrence can keep running while prediction fails



- Mamba-2 was trained with a 2,048-token context.
- Some base checkpoints deteriorate sharply after that range.
- Dashed curves: 500 steps of State Passing (SP) post-training.

Computational reach does not ensure statistical generalization.

Paper, Fig. 1(b), PDF p. 2. Position-wise perplexity: lower is better.

Position-wise perplexity separates early and late failure

$$\mathrm{PPL}(t) = \exp(\mathbb{E}_{x\sim\mathcal{D}}[-\log q(x_t \mid x_{0:t-1})])$$

- Average the negative log-likelihood at a fixed position, then exponentiate.
- Let $p^\star$ be the best PPL in the training range, attained at $t^\star$.

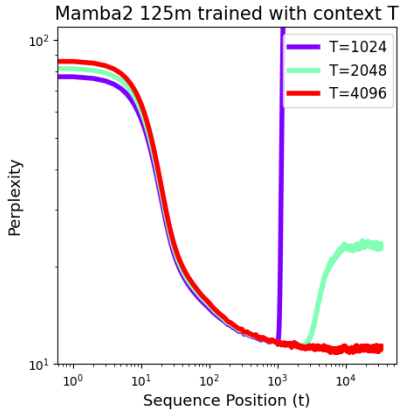
The paper's formal definition

Length generalization to $T > T_{\text{train}}$ requires $\mathrm{PPL}(t) \le p^\star$ for every $t \in [t^\star, T]$.

Empirical plots show approximate trends. Stable PPL alone does not establish use of distant facts.

Paper, Definitions 3.1–3.2, PDF p. 4; motivation for task evaluation, Sec. 5, p. 9.

Longer training contexts improve extrapolation



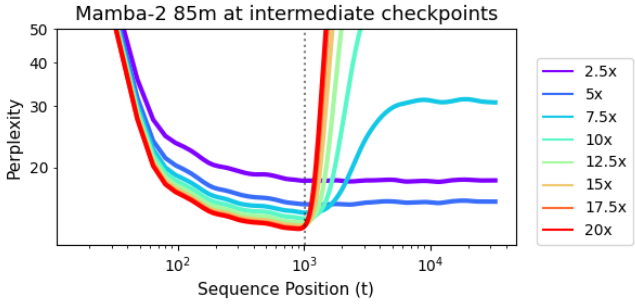
Mamba-2 125M, trained from scratch

- Same model size; 25B training tokens.
- Training contexts: 1,024, 2,048, 4,096.
- The 4,096-token model is substantially more stable at long positions.

Architecture and state size alone do not determine the failure.

Paper, Sec. 3.2 and Fig. 2(c), PDF p. 4. Other panels cover 45M, 85M, and GLA 70M.

More short-context training can worsen extrapolation



- Mamba-2 85M, context 1,024; checkpoints span 4.25B–34B tokens.
- In-range PPL improves as late-position PPL deteriorates.
- This is consistent with overfitting to early-state distributions.

Paper, Sec. 3.3 and Fig. 3, PDF pp. 4–5. Multipliers refer to the paper's Chinchilla budget convention.

Effective Remembrance measures prefix influence

$$\mathrm{EffRem}_T(t) = d(q(\cdot \mid x_{0:T}), q(\cdot \mid x_{t:T}))$$

$$d(p, q) = \mathrm{TV}(p, q) = \frac{1}{2} \sum_{v \in \mathcal{V}} |p(v) - q(v)|$$

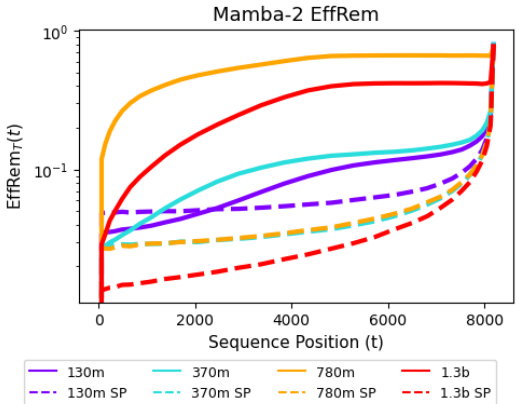
- Fix the prediction position T; remove the prefix $[0, t)$.
- Recompute the retained suffix from the normal initial state.
- High TV means the prefix strongly changes the output distribution.

Interpretation

Sensitivity can reflect useful information or harmful interference. It is not a memory-capacity score.

Paper, Definition 3.3 and Eqs. (2)–(3), PDF p. 5. TV is written in its intended standard form.

Failing models are unusually sensitive to early tokens

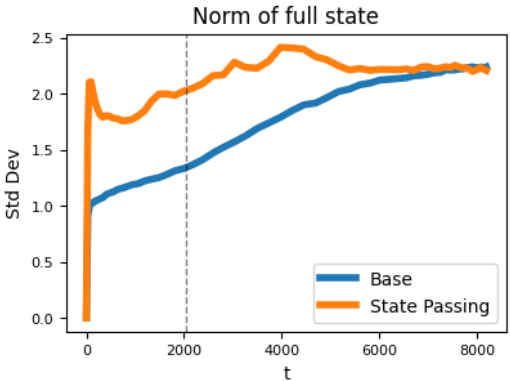


Mamba-2; fixed $T = 8192$

- Solid: base checkpoints.
- Dashed: SP post-training.
- Removing a short, distant prefix can strongly alter a failing model's output.
- SP reduces this dependence while improving long-position PPL.

Paper, Fig. 1(e), PDF p. 2; Sec. 3.4, p. 5. The vertical scale is logarithmic.

State drift matters more than a small state scale



Mamba-2 130M

- Base statistics keep changing beyond training length 2,048.
- SP reaches a more stable range earlier.
- SP often has *larger* values, so simply shrinking states is insufficient.

Paper, Sec. 3.5 and Fig. 4, PDF p. 6. The title says "norm"; the axis says "Std Dev." We interpret the plotted scale trend.

The hypothesis is a mismatch in attainable states

$$\mathcal{D}_t = \mathrm{Law}(h_t(X_{0:t}))$$

Zero-initialized short training

The model sees early distributions $\mathcal{D}_0, \ldots, \mathcal{D}_{T_{\mathrm{train}}-1}$ repeatedly.

Long-sequence evaluation

The same recurrence reaches later distributions that may be unfamiliar.

- With approximately stationary inputs, state shift becomes a useful explanation.
- Prediction: expose training to more realistic states, even in short chunks.

Paper, Sec. 3.6, PDF p. 6. This is a supported hypothesis, not a general error-bound theorem.

A nonzero initial state adds a propagated contribution

$$h_t = A_t h_{t-1} + B_t x_t, \qquad y_t = C_t^\top h_t$$

$$h_t = \underbrace{\sum_{s=0}^{t} A^{\times}_{s+1:t} B_s x_s}_{\text{zero-initialized result}} + \underbrace{A^{\times}_{0:t} h_{-1}}_{\text{initial-state contribution}}$$

$$y_t = y_t^{(0)} + C_t^\top A^{\times}_{0:t} h_{-1}$$

- $h_t \in \mathbb{R}^N$, $A_t \in \mathbb{R}^{N\times N}$, $B_t, C_t \in \mathbb{R}^{N\times 1}$.
- For general matrices, $A^{\times}_{s:t} = A_t \cdots A_s$ is ordered.
- Initial states can be incorporated into parallel recurrence computation.

Paper, Eqs. (1), (4), PDF pp. 3, 7. Ordered-product clarification for the general matrix case.

Four interventions preserve different state structure

Method	Initial state	What it preserves
Random Noise	IID $\mathcal{N}(0, \sigma^2)$	A global noise scale
Fitted Noise	Gaussian samples fitted per layer and head	Per-head mean and variance
State Passing	Final state of a shuffled sequence	Actual joint state structure
TBTT	Final state from the preceding chunk of the same sequence	State + sequence order

Test whether realistic state coverage is more useful than arbitrary noise.

Paper, Secs. 4.2–4.5, PDF pp. 7–9. TBTT: truncated backpropagation through time.

Fitted Noise matches moments within each head

$$\mu^{(li)} \leftarrow (1-\beta)\,\mathrm{Mean}(h_T^{(li)}) + \beta\mu^{(li)},$$
$$\sigma^{2(li)} \leftarrow (1-\beta)\,\mathrm{Var}(h_T^{(li)}) + \beta\sigma^{2(li)},$$
$$h_{-1,j}^{(li)} \overset{\text{independent}}{\sim} \mathcal{N}\left(\mu^{(li)}, \sigma^{2(li)}\right).$$

- l: layer; i: head; j: an element of that head's state.
- Statistics aggregate head and state-expansion dimensions.
- $\beta = 0.1$: **90% current statistics, 10% previous estimate.**
- Independent sampling discards dependencies between state elements.

Paper, Sec. 4.3 and Eqs. (5)–(6), PDF p. 8. Noise is used during training, with zero initialization at validation.

State Passing changes training starts, not test-time inputs

Training data flow

Previous shuffled batch $\longrightarrow$ final states $\longrightarrow$ stop gradient $\longrightarrow$ next batch's initial states.

- Reset the initial state to zero with probability $p = 0.1$.
- Carry forward state history; optimize using the current short chunk.
- TBTT instead preserves the order of chunks from the same long sequence.
- At inference, start a new stream normally and continue the recurrence.

SP does not require semantically adjacent training chunks.

Paper, Secs. 4.4–4.5, PDF pp. 8–9; chunk-order distinction in footnote 3, p. 10. Flow is a teaching schematic.

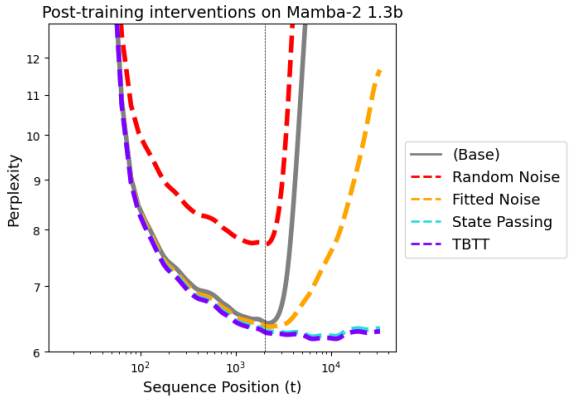
The 32k and 128k results use different budgets

Experiment	Adaptation budget	Context and evaluation
Four-way comparison (Fig. 5)	100 steps	Mamba-2: 2k training; 32k test
Broad evaluation (Fig. 1)	500 steps	Mamba-2 / GLA; up to 128k
Mamba-1 (Fig. 1)	1000 steps	2k training; up to 128k
Passkey (Fig. 7)	1000 task steps	2k training; up to 256k

Pile post-training uses 1/10 of the base peak learning rate.
The reported 100 / 500-step budgets are approximately 0.02% / 0.1% of pre-training.

Paper, Figs. 1, 5, PDF pp. 2, 7; Appendix B, pp. 16–17. GLA's training context is 512, not 2k.

Real states succeed where independent noise fails

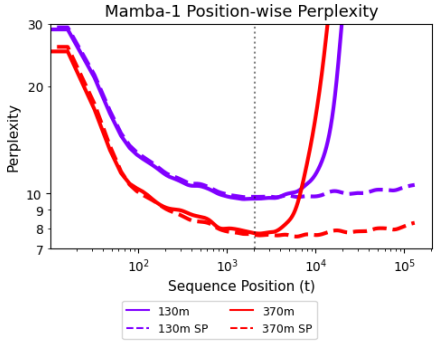


Paper, Fig. 5(c), PDF p. 7; interpretation in Secs. 4.2–4.5, pp. 7–9.

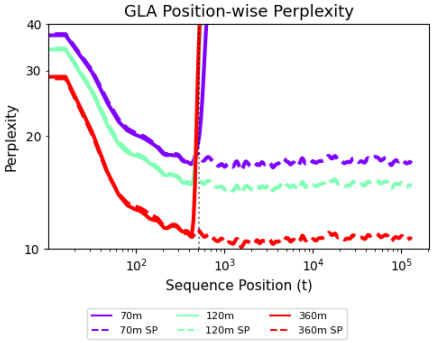
100-step intervention

- Fitted Noise helps 370M and 780M, but fails at 1.3B.
- SP and TBTT remain stable for all three sizes through 32k.
- SP and TBTT are very close despite different sequence continuity.

State Passing also transfers to Mamba-1 and GLA



Mamba-1: 2,048-token context; 1,000 SP steps

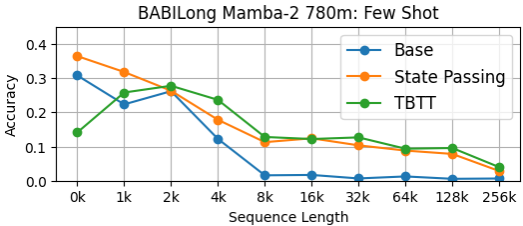


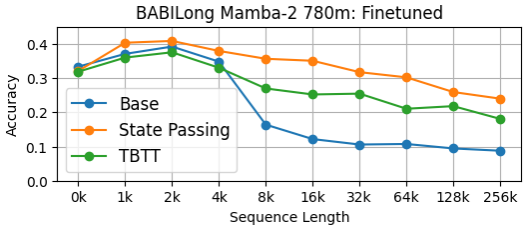
GLA: 512-token context; 500 SP steps

A shared training intervention improves long-position behavior across distinct recurrent mechanisms.

Paper, Fig. 1(a,c), PDF p. 2. Dashed lines denote SP; solid lines denote base models. RWKV-v6 is in backup.

BABILong improves, with substantial task error remaining

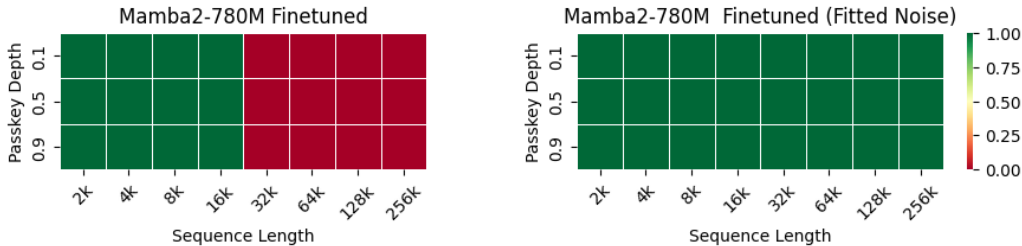




- Mamba-2 780M; tasks qa1–qa10; training / finetuning context 2,048.
- Task finetuning uses one epoch *without* SP or TBTT for every model.
- Earlier Pile post-training benefits survive later ordinary task finetuning.
- Few-shot accuracy at the longest inputs remains very low.

Paper, Sec. 5.1 and Fig. 6, PDF p. 9; Appendix B.3, p. 17. Accuracy is higher-is-better.

Fitted Noise enables 256k passkey retrieval at 780M



- Same task: 1,000 finetuning steps, sequences of length 2,048.
- Fitted Noise reaches 100% on the reported grid, including 256k.
- A terminal passkey state is a poor start for another sample: task phases are nonstationary.

Paper, Sec. 5.2 and Fig. 7(e,f), PDF p. 10; Appendix B.4, p. 17. Horizontal: length; vertical: depth; green: high accuracy.

Copying improves by 20 percentage points, yet stays hard

Training method	Character accuracy ↑
Mamba-2	0.27 ± 0.03
Mamba-2 + State Passing	0.47 ± 0.11

- 45M parameters; trained from scratch on one million sequences.
- Training lengths 50–100; evaluation length 300.
- Length robustness improves; arbitrary-sequence storage remains difficult.

Paper, Sec. 5.3 and Table 1, PDF p. 10. The paper does not define the statistical meaning of the reported $\pm$ values.

State coverage is a useful explanation with concrete limits

- **Mechanism:** realistic initial states improve both predictions and state dynamics; the experiments do not isolate these two effects.
- **Task dependence:** nonstationary task phases can make direct state reuse inappropriate.
- **Capacity:** success on one passkey does not imply reliable copying of arbitrary long sequences.
- **Scope:** evidence covers Mamba-1/2, GLA, and RWKV-v6; hybrid architectures and other domains remain open.

Synthesis of Paper, Secs. 3–5 and 7, PDF pp. 4–11; Appendix B.4, p. 17. Mechanism separation is our analysis.

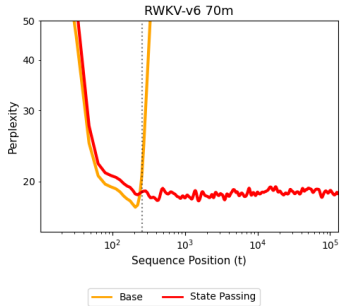
Train on the states that long sequences will produce

1. Diagnose prediction stability, prefix influence, and distant-information use separately.
2. Treat the initial-state distribution as part of the training recipe.
3. Test inexpensive state interventions before attributing failure solely to architecture or capacity.

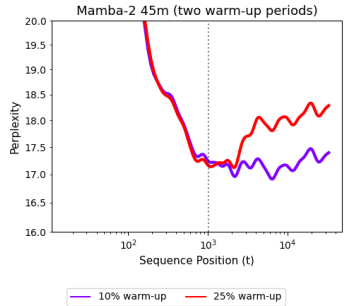
Novelty in context: SP/TBTT existed in prior work. This paper contributes the state-distribution account, Effective Remembrance, intervention comparisons, and efficient post-training evidence.

Ricardo Buitrago Ruiz and Albert Gu. Understanding and Improving Length Generalization in Recurrent Models. arXiv:2507.02782v2 (2025). Versioned paper. Related-work attribution: Sec. 6, PDF p. 11.

Backup: RWKV-v6 and warm-up support the broader picture



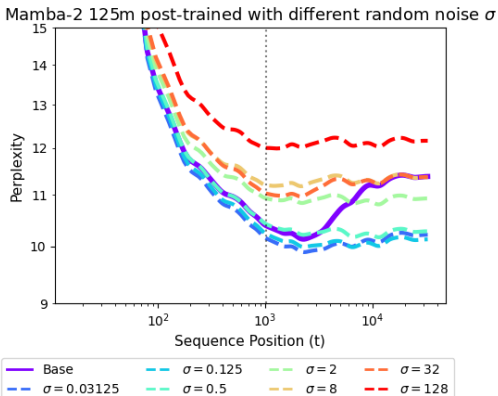
RWKV-v6 70M, context 256.
35B pretraining tokens; 250M SP tokens.



Mamba-2 45M, context 1,024, 9B tokens.
25% warm-up hurts extrapolation relative to 10%.

Paper, Appendices D–E, Figs. 10–11, PDF pp. 17–19. These are separate experiments with separate budgets.

Backup: arbitrary noise requires a suitable scale

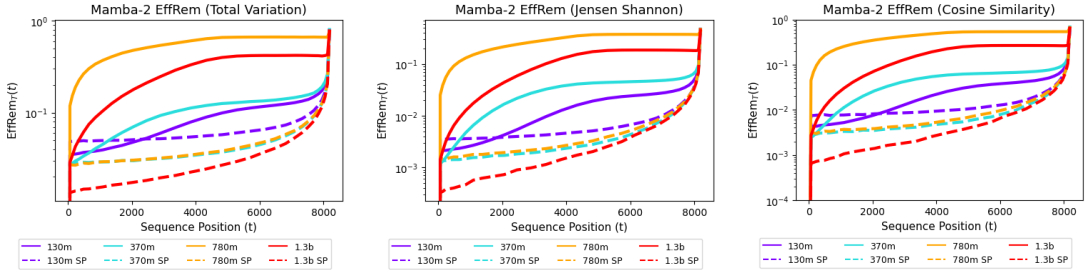


Mamba-2 125M; 100 steps

- Initial state elements are IID $\mathcal{N}(0, \sigma^2)$.
- Too little noise may fail to fix generalization.
- Too much noise worsens PPL.
- Validation still starts from zero.

Paper, Appendix F and Fig. 12, PDF pp. 19–20; training/validation distinction in Sec. 4.2, p. 8.

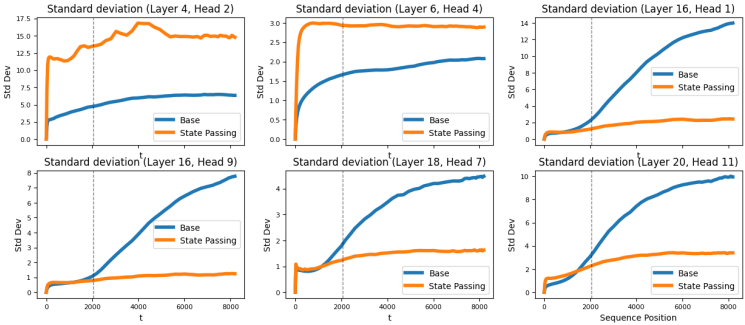
Backup: output-distance choices preserve the main pattern



- Mamba-2 base and SP models, fixed prediction position $T = 8192$.
- Absolute values differ, but the characteristic curves are similar.
- TV is bounded by 1; reaching 1 is not a mathematical necessity.

Paper, Appendix G and Fig. 13, PDF p. 20. The paper notes that 1−cosine similarity is not a proper metric.

Backup: selected heads reveal where the state drift occurs



- Heads were selected for pronounced variation; most heads are less abnormal.
- SP can have larger standard deviation and still be more stable over time.

Paper, Appendix C and Fig. 9, PDF p. 18. These moments do not fully characterize the joint state distribution.

Backup: preserve matrix order and parameter conventions

$$\Phi(t,s) = A_t A_{t-1} \cdots A_s, \qquad h_t = \Phi(t,0)h_{-1} + \sum_{s=0}^{t} \Phi(t,s+1)B_s x_s$$

- General matrices need ordered products; Appendix H swaps factors in an intermediate line.
- Appendix I assumes diagonal transitions. Its `exp(cumsum(A))` requires log-transition values, inferred from the operation.
- The function's output `y` is only the initial-state contribution. Add it to the zero-state output.

Paper, Appendices H–I and Fig. 14, PDF pp. 20–21. Parameter-convention and ordering clarifications are our analysis.