Understanding and Improving Length Generalization in Recurrent Models

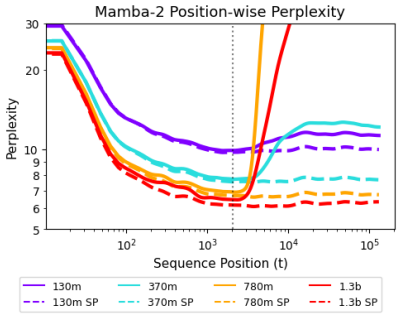
Research discussion – 20–30 minute main presentation

Paper authors: Ricardo Buitrago Ruiz and Albert Gu

arXiv:2507.02782v2 – Revised October 12, 2025
Reading prepared October 13, 2026

Broaden the states seen during training to improve long-sequence behavior.

A recurrent update can run longer than it works well



$$h_t = A_t h_{t-1} + B_t x_t$$

- Fixed-size state; linear sequence cost.
- Mamba-2 training context: 2048.
- Solid: baseline; dashed: State Passing (SP).
- The same architecture can behave very differently beyond training.

Paper, Fig. 1(b), PDF p. 2; Eq. (1), p. 3. Lower perplexity is better.

Length generalization concerns prediction quality

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

Average negative log-likelihood at a fixed position, then exponentiate.

The paper's definition

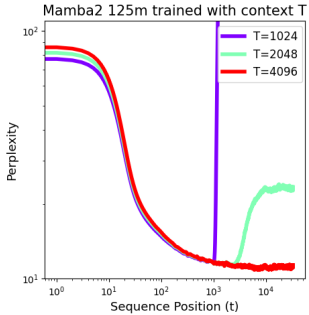
Let $p^\star$ be the minimum training-range PPL, reached at $t^\star$. Generalization to $T > T_{\text{train}}$ requires

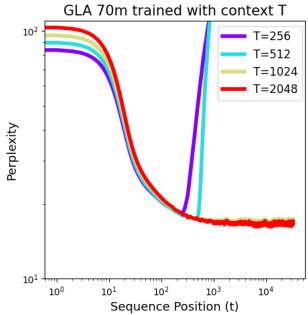
$$\mathrm{PPL}(t) \le p^\star \quad \text{for all } t \in [t^\star, T].$$

Stable PPL still leaves open whether the model uses distant information.

Paper, Definitions 3.1–3.2, PDF p. 4; motivation for long-context tasks, p. 9.

Longer training contexts reduce extrapolation failure

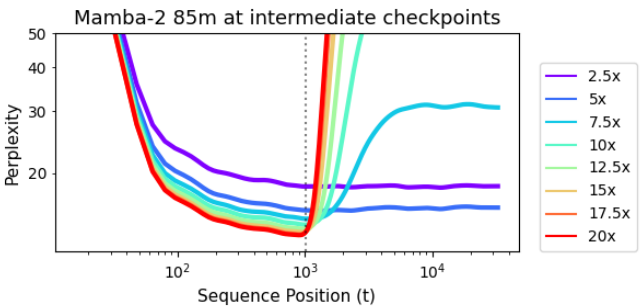




Same model within each panel; colors vary training context length. Longer contexts expose the model to more of its state trajectories.

Paper, Fig. 2(c–d), PDF p. 4. Mamba-2 125M: 25B tokens; GLA 70M: 34B tokens.

More training can improve local fit and harm extrapolation



Mamba-2 85M, context 1024.

- Checkpoints: 4.25B–34B tokens.
- Legend: multiples of the Chinchilla allocation.
- Later checkpoints fit short contexts better but degrade farther out.

Paper, Sec. 3.3 and Fig. 3, PDF pp. 4–5. Warm-up also affects extrapolation (backup).

Effective Remembrance probes sensitivity to an old prefix

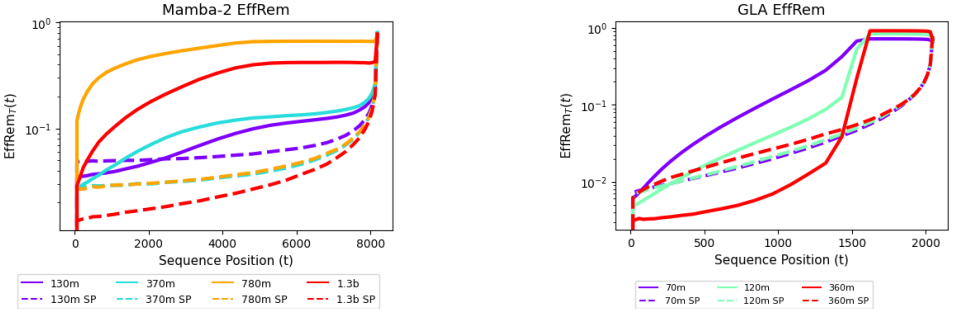
$$\mathrm{EffRem}_T(t) = \mathrm{TV}\big(q(\cdot \mid x_{0:T}), q(\cdot \mid x_{t:T})\big)\,,$$

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

- Fix prediction position T; restart on suffix $x_{t:T}$.
- Large distance at small t: an old prefix strongly affects the output.
- Task accuracy tells us whether the influence is useful.

Paper, Definition 3.3 and Eqs. (2)–(3), PDF p. 5. TV expression corrects a prime-placement typo.

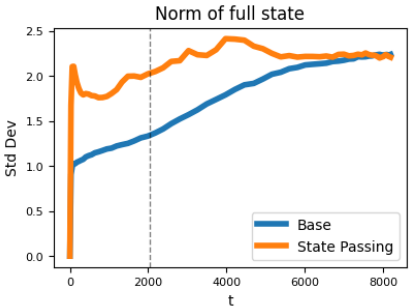
State Passing reduces distant-prefix sensitivity on text



Solid: baseline. Dashed: SP. Colors identify model sizes. Mamba-2 uses $T = 8192$; GLA uses $T = 2048$.

Paper, Fig. 1(e–f), PDF p. 2; Sec. 3.4, p. 5. Vertical axes are logarithmic.

The hypothesis concerns which states training covers



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

Short training exposes only early $\mathcal{D}_t$. Long evaluation can reach unfamiliar distributions.

SP produces more stable state statistics, even when their magnitude is larger.

Paper, Secs. 3.5–3.6 and Fig. 4, PDF p. 6. Plot mixes a Norm title with a Std Dev axis.

A nonzero initial state adds a propagated contribution

For a given coefficient sequence, unroll the linear recurrence:

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

$$y_t = C_t^{\top} \sum_{s=0}^{t} A^{\times}_{s+1:t} B_s x_s + C_t^{\top} A^{\times}_{0:t} h_{-1}.$$

$$A^{\times}_{s:t} = A_t A_{t-1} \cdots A_s, \qquad h_t, B_t, C_t \in \mathbb{R}^N, \quad A_t \in \mathbb{R}^{N \times N}.$$

Scan or chunk computations can incorporate the extra initial-state term.

Paper, Eq. (4), PDF p. 7; Appendix H, p. 20. Ordered products are required for general matrices.

Four interventions vary the initial-state distribution

Intervention	Initial state	Structure retained
Random Noise	IID $\mathcal{N}(0, \sigma^2)$	One global scale
Fitted Noise	Gaussian fitted per layer/head	Means and variances
State Passing	Final state of another sequence	Full generated state
TBTT	Final state of preceding chunk	State and input continuity

The comparison tests state coverage as well as context continuity.

Paper, Secs. 4.2–4.5, PDF pp. 7–9.

State Passing carries states between shuffled batches

Training flow (teaching schematic)

Shuffled batch $k \longrightarrow$ final states $\longrightarrow$ detach $\longrightarrow$ initial states of batch $k + 1$.

- Keep the original short sequence length for each training example.
- Reset the initial state to zero with probability $p = 0.1$.
- Maintain state tensors for the model's layers and heads.
- TBTT instead preserves consecutive chunks from one long sequence.

After training, use the learned weights with ordinary recurrent inference.

Paper, Secs. 4.4–4.5, PDF pp. 8–9. Schematic reorganizes the described procedure.

Fitted Noise matches statistics by layer and head

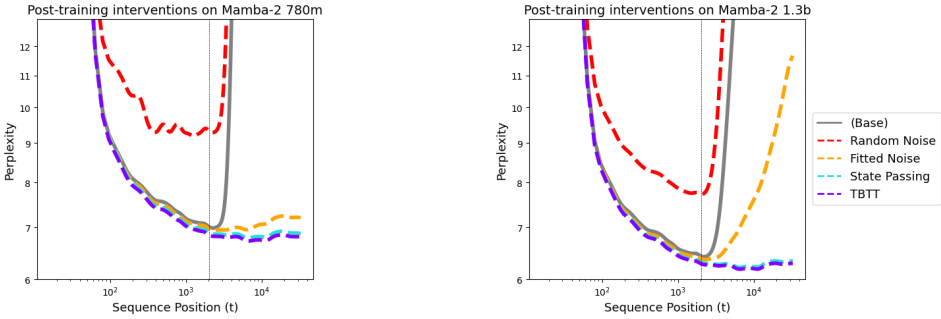
For layer l, head i, update final-state statistics:

$$\mu^{(li)} \leftarrow (1-\beta)\,\mathrm{Mean}(h_T^{(li)}) + \beta\mu^{(li)},$$
$$(\sigma^2)^{(li)} \leftarrow (1-\beta)\,\mathrm{Var}(h_T^{(li)}) + \beta(\sigma^2)^{(li)}.$$

- $\beta = 0.1$: new statistics receive weight 0.9.
- Sample entries independently from $\mathcal{N}(\mu^{(li)}, (\sigma^2)^{(li)})$.
- Statistics cover head channels and state expansion dimensions.
- Noise is used in training; validation starts from zero.

Paper, Secs. 4.2–4.3 and Eqs. (5)–(6), PDF pp. 7–8.

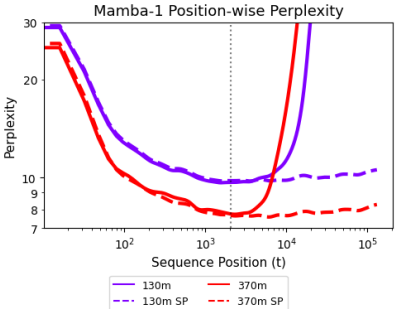
At 100 steps, SP and TBTT are the most consistent

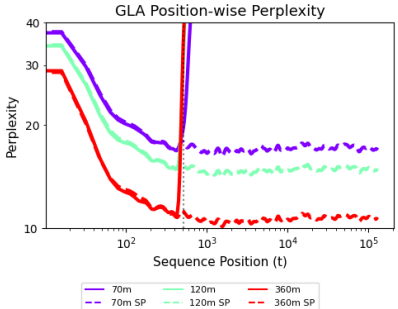


Fitted Noise helps 370M and 780M, but fails at long positions for 1.3B. SP and TBTT remain effective through 32K in all three tested sizes.

Paper, Fig. 5 (780M and 1.3B shown), PDF p. 7. All interventions: 100 post-training steps.

State Passing also improves Mamba-1 and GLA





Mamba-1: $T_{\text{train}} = 2048$, 1000 SP steps. GLA: $T_{\text{train}} = 512$, 500 SP steps.

Paper, Fig. 1(a,c), PDF p. 2. Tests extend to roughly 128K; curves need not be perfectly flat.

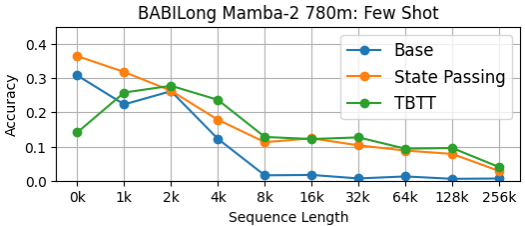
The headline results use several distinct budgets

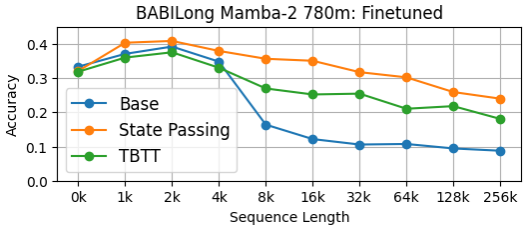
Experiment	Train context	Added training	Evaluation
Mamba-2, four methods	2K	100 steps	32K PPL
Mamba-2, SP	2K	500 steps	$\sim$ 128K PPL
Mamba-1, SP	2K	1000 steps	$\sim$ 128K PPL
GLA, SP	512	500 steps	$\sim$ 128K PPL
Passkey fine-tuning	2K	1000 steps	Up to 256K

Pile interventions use a peak learning rate one tenth of the tabulated base recipe.

Paper, Figs. 1 and 5; Appendix B.2/B.4, PDF p. 17. Percent-budget headlines are setting-specific.

BABILong improves, while accuracy still falls with length

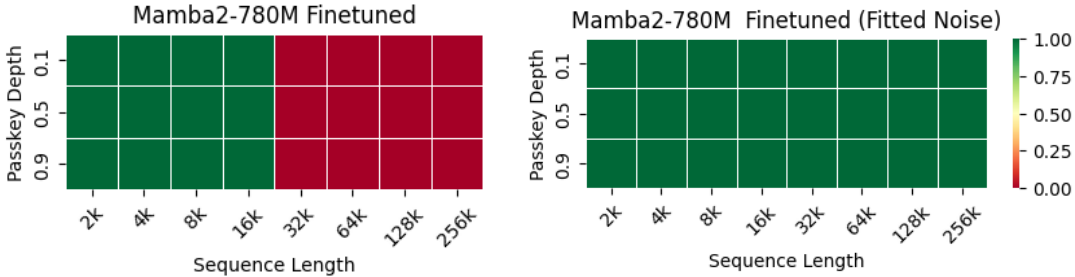




Mamba-2 780M; qa1–qa10; 2K training/fine-tuning context. Task fine-tuning is one epoch *without* SP or TBTT in every group.

Paper, Fig. 6, PDF p. 9; Appendix B.3, p. 17. At 256K, fine-tuned SP is about 0.24 (visual estimate).

Fitted Noise enables 256K passkey retrieval for 780M



- Both are fine-tuned for 1000 steps on 2K sequences.
- Green denotes accuracy 1; rows vary passkey depth.
- Fitted Noise avoids carrying an answer-bearing terminal state into a new task.

Paper, Fig. 7(e–f) and Sec. 5.2, PDF p. 10; Appendix B.4, p. 17. Task-specific fine-tuning.

Copying improves by 20 percentage points, but remains hard

Training method	Character accuracy at length 300
Mamba-2	0.27 ± 0.03
Mamba-2 + State Passing	0.47 ± 0.11

- 45M parameters, trained from random initialization.
- One million sequences; training lengths 50–100.
- Character accuracy differs from exact whole-sequence copying.

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

The contribution is a state-distribution explanation

Prior work acknowledged by the authors

LongSSM already examined zero initialization and proposed State Passing / TBTT pre-training.

This paper adds:

- The unexplored-states hypothesis and an output-level diagnostic.
- A controlled comparison of initial-state distributions.
- Low-budget post-training across several recurrent architectures.
- Evidence on reasoning, retrieval, and copying.

Paper, Sec. 6, PDF p. 11; Appendix A, p. 16. Prior-work discussion follows the paper's attribution.

Practical choices depend on the state and input together

- **State Passing:** retains a full generated state; works well with shuffled text.
- **TBTT:** retains real temporal continuity; needs ordered chunks.
- **Fitted Noise:** controls the initial distribution; omits correlations.

What changes the interpretation

Stable low-order statistics do not identify a full state distribution. Passkey inputs are nonstationary, so an actual terminal state may be a poor initial state.

Fig. 5 uses the original checkpoint as Base; no matched 100-step zero-state continuation control is shown.

Paper, Secs. 3.6 and 4, pp. 6–9; Appendix B.4, p. 17. Joint-distribution framing is reading analysis.

Train the model on states it will encounter later

1. Short, zero-initialized training can leave long-run states unexplored.
2. A small amount of training with realistic initial states can repair severe extrapolation failure.
3. Check distant-information tasks as well as PPL: retrieval can succeed while copying and reasoning remain difficult.

Initial-state distributions are a useful training design variable.

Synthesis of the results. Other distributions, S4, and hybrid models are future directions (Sec. 7, p. 11).

Reference and backup material

Ricardo Buitrago Ruiz and Albert Gu.

Understanding and Improving Length Generalization in Recurrent Models.

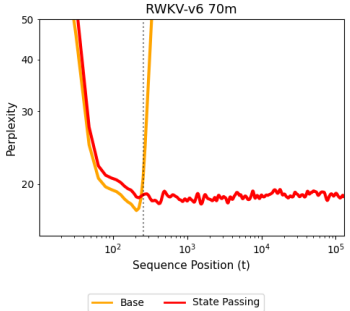
arXiv:2507.02782v2, October 12, 2025.

https://arxiv.org/abs/2507.02782v2

- Experimental plots are author figures from this version.
- Tables reproduce selected results or reorganize reported settings.
- Teaching examples and interpretation are identified where used.

Prepared as an independent research discussion. All page references refer to the paper PDF.

Backup: RWKV-v6 provides another architecture test

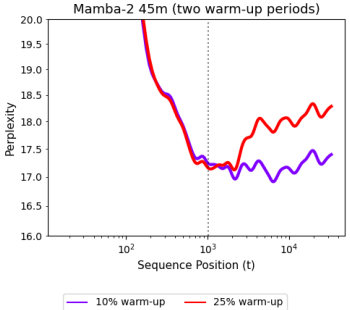


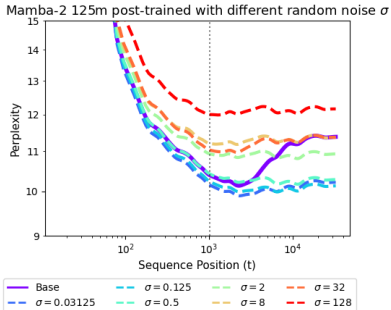
- 70M parameters, context 256.
- Pre-training: 35B tokens.
- SP post-training: 250M tokens.
- Baseline diverges; SP stays much more stable.

Added token budget: about 0.71% in this experiment.

Paper, Appendix D and Fig. 10, PDF pp. 17–18. Percentage computed from the token counts.

Backup: warm-up and noise scale both matter

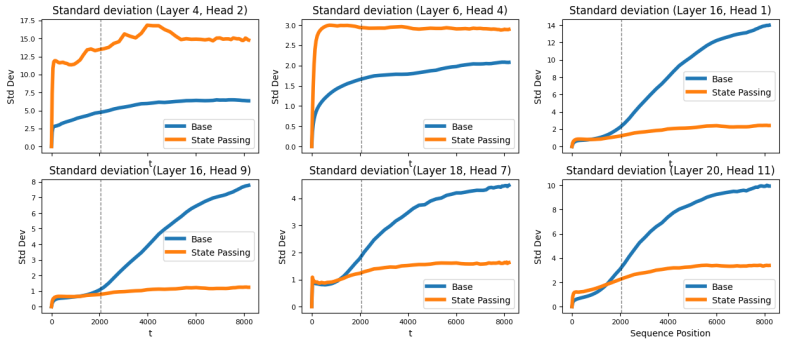




Longer warm-up worsens extrapolation in the 45M example. Excessive initial-state noise hurts PPL in the 125M example.

Paper, Figs. 11–12, PDF p. 19. Left: 9B tokens, context 1024. Right: 100 post-training steps.

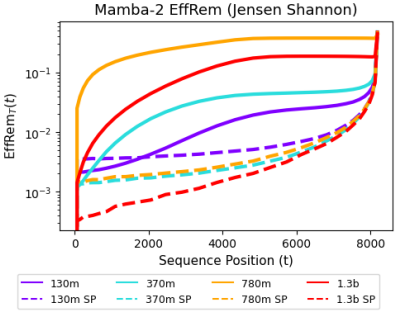
Backup: a few heads explain much of the state shift

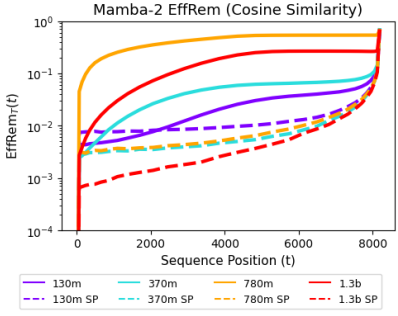


Heads were selected for strong variation. SP sometimes increases standard deviation while stabilizing its long-position behavior.

Paper, Fig. 9 and Appendix C, PDF pp. 17–18. Mamba-2 130M; most other heads lack this abnormal growth.

Backup: the sensitivity pattern survives other distances





Shapes are similar to TV, but absolute values differ. The right panel uses 1−cosine similarity despite its abbreviated title.

Paper, Fig. 13(b–c) and Appendix G, PDF p. 20. Jensen–Shannon distance/divergence naming varies.

Backup: implement the transition product with care

For diagonal positive transitions, let a_t be their entries and $\ell_t = \log a_t$.

$$A^{\times}_{0:t} h_{-1} = \exp\left(\sum_{u=0}^{t} \ell_u\right) \odot h_{-1}.$$

- A prefix sum in log space implements the elementwise product.
- Add the initial-state contribution to the zero-state output.
- The paper uses $\exp(\mathrm{cumsum}(A))$: interpret its code variable in log-transition space.
- General matrices require ordered products $A_t \cdots A_s$.

Paper, Appendices H–I and Fig. 14, PDF pp. 20–21. Clarification from the recurrence and code.

Backup: a stable recurrence can have an unseen state scale

Teaching example, not a paper experiment:

$$h_t = a h_{t-1} + x_t, \quad h_{-1} = 0, \quad |a| < 1,$$

with IID, zero-mean, unit-variance inputs.

$$\mathrm{Var}(h_t) = \frac{1 - a^{2(t+1)}}{1 - a^2}, \qquad \lim_{t \to \infty} \mathrm{Var}(h_t) = \frac{1}{1 - a^2}.$$

For $a = 0.99$:

- After 32 tokens: 47.4% of the limiting variance.
- After 512 tokens: approximately the limiting variance.

The state variance stays finite, yet changes beyond a short training horizon.

Illustrative derivation for the state-distribution idea in Sec. 3.6, PDF p. 6.