# Samba: Simple Hybrid State Space Models

## For Efficient Unlimited Context Language Modeling

Paper authors: Liliang Ren, Yang Liu, Yadong Lu,
Yelong Shen, Chen Liang, Weizhu Chen

ICLR 2025 | arXiv:2406.07522v3 (28 Feb 2025)
Research discussion | 16 Oct 2026

Recurrent compression + precise local access, with a fixed history cache.

# The problem is how to represent history

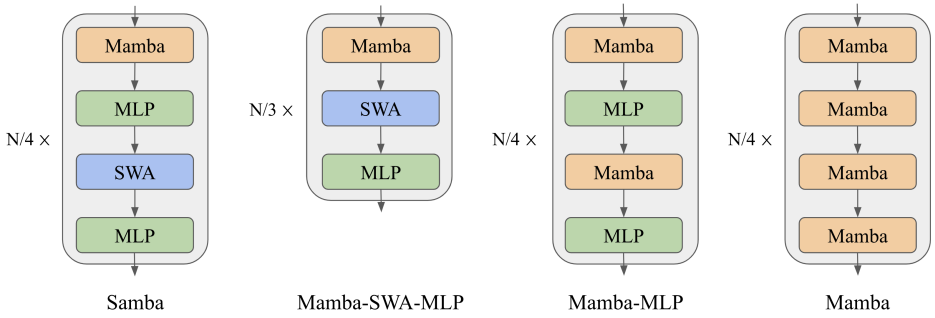
| Mechanism | History representation | Main trade-off |
|---|---|---|
| Full attention | Explicit KV for all past tokens | Direct access; growing cache and work |
| Mamba | Fixed-dimensional recurrent state | Efficient streaming; compressed detail |
| Sliding window attention | Explicit KV for recent tokens | Precise local access; bounded direct reach |

## Samba's question

Can these memory mechanisms complement each other without a global attention layer?

Paper, Secs. 1–2, PDF pp. 1–4.

# Samba alternates two mixers, each followed by an MLP




**One repeat: Mamba → MLP → SWA → MLP**

1.7B Samba: 12 Mamba + 12 SWA + 24 MLP sublayers.
Each sublayer has Pre-Norm and a residual connection.

Paper, Fig. 1, PDF p. 3; Table 12, PDF p. 24. Original author figure.

# Mamba makes retention and writing input-dependent

$$\mathbf{U} = \mathrm{SiLU}(\mathrm{DepthwiseConv}(\mathbf{X}W_{\mathrm{in}})), \qquad \Delta = \mathrm{Softplus}(\mathbf{U}W_r W_q + \mathbf{b})$$

$$\boxed{Z_{t,i,j} = e^{-\Delta_{t,i} e^{A_{i,j}}} Z_{t-1,i,j} + \Delta_{t,i} U_{t,i} B_{t,j}}$$

$$Y_{t,i} = \sum_j Z_{t,i,j} C_{t,j} + D_i U_{t,i}, \quad \mathbf{B} = \mathbf{U}W_b, \quad \mathbf{C} = \mathbf{U}W_c$$

- $i$: expanded channel; $j$: state coordinate; $\mathbf{Z}_t \in \mathbb{R}^{d_e \times d_s}$.
- $d_e = 2d_m$, $d_s = 16$; short convolution width $k = 4$.
- Larger $\Delta$: stronger decay and a larger current-input write factor.

Paper, Sec. 2.1.1, Eq. (1)–(2) and unnumbered recurrence, PDF p. 3. Elementwise expansion for clarity.

# Local attention preserves detail that compression can lose

$$\mathcal{W}_t = \{\max(1, t - w + 1), \ldots, t\}, \qquad w = 2048$$

$$o_t = \sum_{j \in \mathcal{W}_t} \text{softmax}_{j \in \mathcal{W}_t} \left( \frac{q_t^\top k_j}{\sqrt{d_h}} \right) v_j$$

- RoPE within the window; base frequency 10,000.
- Direct access to recent KV; older information travels through states and layers.
- FlashAttention 2 for SWA; a separate SwiGLU MLP follows each mixer.

**Design intuition**

Let attention handle precise local access, while recurrence carries information forward.

Paper, Secs. 2.1.2–2.1.3, PDF p. 4. SWA equation is a teaching restatement; $q, k$ include RoPE.

# A fixed history cache permits continuing the stream

| Mixer | Mixing work for $n$ tokens | History stored per layer |
|---|---|---|
| Full attention | $O(n^2 d_m)$ | $O(n h_{kv} d_h)$ |
| SWA | $O(nwd_m)$ | $O(w h_{kv} d_h)$ |
| Mamba update | $O(n d_e d_s)$ | $O(d_e d_s + k d_e)$ |

- Fix $w$, state size, width, and depth: total sequence work is linear in $n$.
- History storage per stream stops growing after the window fills.
- Finite state capacity still constrains what can be recalled exactly.

Complexity explanation from Sec. 2, PDF pp. 2–4; common projections omitted. Training activations are separate.

# Four regimes test different questions

| Samba size | Data | Training tokens | Main purpose |
|---|---|---|---|
| 421M | SlimPajama | 20B | Architecture ablations |
| 1.3B | SlimPajama | 100B | Scaling and training speed |
| 1.7B | Phi-2 | 230B | Downstream tasks, long context |
| 3.8B | Phi-3 | 3.2T | Large-scale quality, post-training |

- All four use 4,096-token training sequences and a 2,048-token SWA window.
- Strongest architecture comparisons match data and recipes: TFM++ / Phi-3.
- Original and June post-training recipes produce distinct Samba models.

Paper, Sec. 3, PDF pp. 4–8; Appendix B; Table 12, PDF p. 24.

# Matched 3.8B pretraining improves the average by 4.1 points

| Base model metric ↑ | TFM++ | Samba | Difference |
|---|---|---|---|
| MMLU | 67.2 | 71.2 | +4.0 |
| HellaSwag | 76.6 | 77.4 | +0.8 |
| ARC-C | 53.8 | 55.7 | +1.9 |
| WinoGrande | 72.6 | 77.1 | +4.5 |
| TruthfulQA MC2 | 47.3 | 43.4 | −3.9 |
| GSM8K | 51.5 | 69.6 | +18.1 |
| HumanEval | 51.8 | 54.9 | +3.1 |
| Average | 60.1 | **64.2** | +4.1 |

Same 3.2T-token data and recipe. GSM8K gain: **18.1 percentage points**.

Paper, Table 7, PDF p. 20. MMLU/GSM8K: 5-shot; others: zero-shot. ARC-C/HellaSwag: character-normalized.

# Post-training improves short tasks and GovReport

| Metric ↑ | Phi-3-mini | Samba-3.8B-IT | Difference |
|---|---|---|---|
| MMLU | 68.8 | 71.9 | +3.1 |
| GSM8K | 82.5 | 87.6 | +5.1 |
| HumanEval | 58.5 | 62.8 | +4.3 |
| GovReport (ROUGE-L) | 14.4 | 18.9 | +4.5 |
| SQuALITY (ROUGE-L) | 21.6 | 21.2 | −0.4 |

- Original post-training recipe; MMLU 5-shot, GSM8K 8-shot CoT.
- Mean document lengths: GovReport 11,533; SQuALITY 7,974 tokens.

Paper, Table 1, PDF p. 4; Appendix F, p. 23. HumanEval: zero-shot pass@1. Phi-3 row marked as report-derived.

# The best hybrid arrangement depends on the task

| Phi-2 model | Avg. (15) | SQuAD | MMLU | GSM8K |
|---|---|---|---|---|
| Llama-3 1.6B | 51.17 | 74.88 | 43.84 | 32.68 |
| Mistral 1.6B | 51.12 | 72.82 | 43.54 | 32.45 |
| Mamba 1.8B | 52.31 | 67.66 | 45.28 | 32.07 |
| Mamba-SWA-MLP 1.6B | 53.77 | 76.73 | 47.39 | **44.05** |
| Mamba-MLP 1.9B | 51.38 | 63.86 | 43.68 | 27.52 |
| Samba 1.7B | **54.33** | **77.64** | **48.01** | 38.97 |

- Samba gains 9.98 SQuAD points over pure Mamba.
- Mamba-SWA-MLP wins on GSM8K by 5.08 percentage points.

Paper, Table 2, PDF p. 5. All trained on 230B Phi-2 tokens; approximate parameter matching. GSM8K: 5-shot.

# SWA and RoPE are critical to the observed extrapolation

| About 1.3B, SlimPajama 100B | PPL @4K ↓ | @8K ↓ | @16K ↓ |
|---|---|---|---|
| Llama-2 | 7.60 | 44.32 | 249.64 |
| Llama-2-SWA | 7.60 | 7.37 | 7.21 |
| Mamba | 7.47 | 7.26 | 7.15 |
| Sliding GLA (1.2B) | 7.58 | 7.35 | 7.19 |
| Sliding RetNet (1.4B) | 7.56 | 7.35 | 7.56 |
| Mamba-SWA-MLP | 7.37 | 7.16 | 7.00 |
| MLP2-SWA-MLP | 7.81 | 7.58 | 7.42 |
| Samba-NoPE | 7.33 | 20.40 | 326.17 |
| Samba | **7.32** | **7.11** | **6.96** |

Removing RoPE barely changes 4K PPL, but destroys length extrapolation.

Paper, Table 3, PDF p. 6. Training length 4,096; SWA window 2,048. Reported PPL fluctuation: about $\pm 0.3\%$.

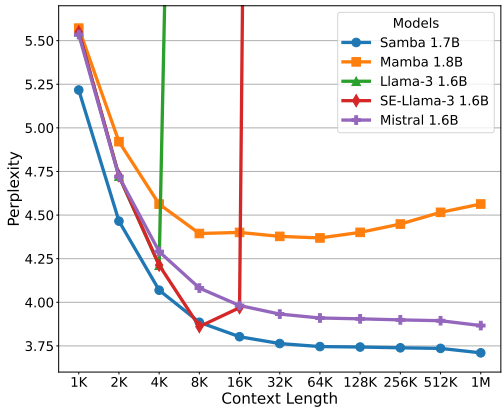
# Fewer recurrent layers can make a hybrid faster to train

| About 1.3B, 4K sequences | Layers | Throughput ($10^5$ tokens/s) |
|---|---|---|
| Mamba | 48 | 17.8 |
| Mamba-SWA-MLP | 36 | 23.5 |
| Samba | 36 | **25.2** |
| Llama-2-SWA | 40 | 26.2 |
| MLP2-SWA-MLP | 36 | 26.6 |

- Samba / Mamba: $25.2/17.8 \approx 1.42\times$ training throughput.
- MLP-heavy designs use efficient dense operations, but removing recurrence hurts PPL.

Paper, Table 3 and Sec. 3.2, PDF pp. 6–7. Measured on 64 H100 GPUs; ratio calculated from the table.

# Proof-Pile prediction improves far beyond 4K training

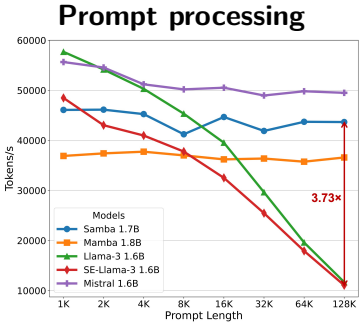



- Samba 1.7B stays stable up to 1M tokens.
- SWA-only Mistral also extrapolates, at higher PPL.
- Pure Mamba deteriorates gradually at the long end.

Prediction quality and exact retrieval are separate tests.

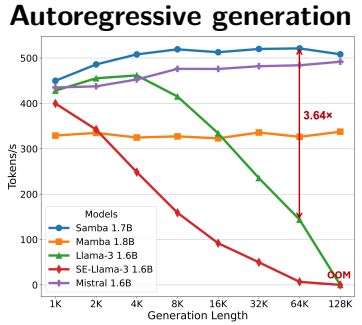
Paper, Fig. 2a and Sec. 3.3, PDF pp. 7–8. Original figure. PPL evaluation uses a 4,096-token sliding-window approach.

# Inference gains grow when full attention must read more history




3.73× at 128K prompt length

Total input tokens fixed at 128K; batch varies.



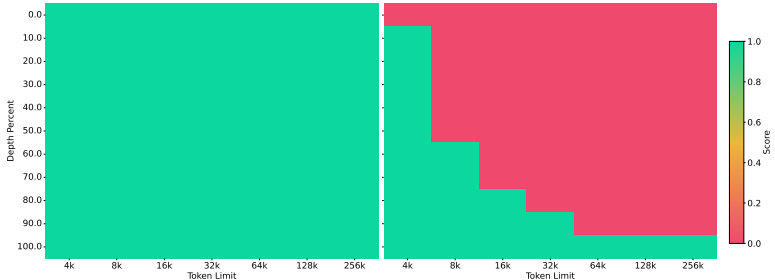

3.64× at 64K generation length

Batch size fixed at 16.

Both ratios compare Samba 1.7B with Llama-3 1.6B.

Paper, Figs. 6 / 2b, PDF pp. 21 / 7; Sec. 3.3, p. 8: single A100, bfloat16, 10 measurements averaged. Original figures.

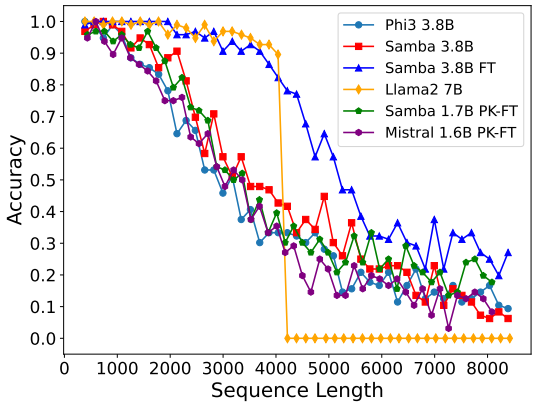
# Task tuning enables 256K Passkey recall from 4K examples




- Samba 1.7B (left) vs. Mistral 1.6B (right); 500 steps at 4K.
- Seven lengths, 11 insertion depths, five keys per cell: 385 tests.
- Samba reaches 100% on the final grid; the base model lacks this behavior.

Paper, Fig. 3, PDF p. 8; Appendix C, p. 22. Batch 2,048: nominal 4.19B input tokens over 500 steps (calculated).

# Many-item retrieval exposes a capacity and transfer gap




- 20–480 name–number pairs; random target query.
- Samba 3.8B: 100 Phonebook tuning steps at 4K improve extrapolation.
- Passkey-tuned 1.7B models transfer poorly to Phonebook.

Paper, Fig. 4 and Sec. 3.4, PDF pp. 8–9. Original figure. Longer inputs also contain more items.

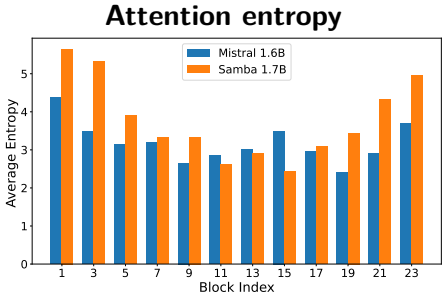
# Full-attention hybrids trail SWA, with an important table nuance

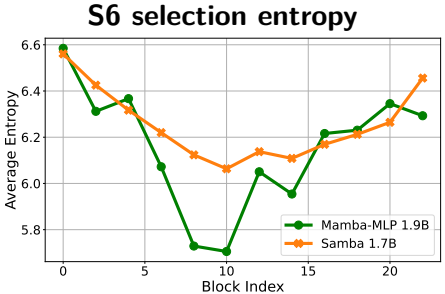
| Full-attention block indices | PPL @4K | @8K | @16K |
|---|---|---|---|
| 11 | 10.29 | 10.53 | 13.66 |
| 5 | 10.10 | 10.05 | 12.83 |
| 0 | 10.89 | 10.55 | 10.63 |
| 1, 5 | 10.06 | 10.34 | 13.57 |
| Samba: SWA at odd indices | 10.06 | 9.65 | **9.57** |

- All reported full-attention variants have worse 16K PPL than Samba.
- The block-0 row does *not* show the "explosion" described in the prose.

Paper, Table 5 and Sec. 4, PDF p. 9. Approximately 430M; 20B SlimPajama tokens; 12 mixer+MLP blocks.

# Entropy patterns suggest specialization across layers




- Samba attention is more concentrated in middle layers, broader at the ends.
- Its middle Mamba layers have more dispersed selection than Mamba-MLP.
- Authors' interpretation: local retrieval frees recurrence for other structure.

Paper, Fig. 5, PDF p. 10; entropy definitions in Appendix E, p. 23. 100 GSM8K samples. Observational evidence.

# Three kinds of long-context evidence must stay separate

| Capability | Evidence | Practical implication |
|---|---|---|
| Streaming computation | Fixed state + 2K KV window | History cache can remain bounded |
| Language prediction | 4K training; Proof-Pile up to 1M | Useful prediction survives long input |
| Exact retrieval | Tuned Passkey to 256K; harder Phonebook | Retention depends on task and memory load |

## Research question

How can ordinary pretraining produce retrieval that transfers across tasks and item counts?

Synthesis of Secs. 2–3 and Appendix H, PDF pp. 2–9, 25.

# Paper and discussion pointers

**Liliang Ren, Yang Liu, Yadong Lu, Yelong Shen, Chen Liang, Weizhu Chen.**

*Samba: Simple Hybrid State Space Models for Efficient Unlimited Context Language Modeling.* ICLR 2025; arXiv:2406.07522v3.

arxiv.org/abs/2406.07522v3
github.com/microsoft/Samba

- Architecture and recurrence: Fig. 1, PDF pp. 2–4.
- Matched-data quality: Tables 1, 2, 7, 8.
- Extrapolation and recall: Figs. 2–4, 6–8.

Figures in this presentation are the authors' originals. Tables select and reformat reported entries.

# Backup: June post-training is a separate comparison

| Chat-mode evaluation | Samba (June) 3.8B | Phi3 (June) 3.8B |
|---|---|---|
| MMLU, 5-shot | 69.0 | 67.2 |
| MMLU-Pro, zero-shot CoT | 47.9 | 46.5 |
| ARC-C, 10-shot | 87.8 | 86.8 |
| GPQA, zero-shot CoT | 29.5 | 29.0 |
| GSM8K, 8-shot CoT | 86.4 | 84.8 |
| HumanEval, zero-shot | 70.1 | 66.5 |
| MBPP, 3-shot | 71.7 | 70.0 |
| Average | **66.1** | 64.4 |

Same data and recipe within this pair; do not merge with the original IT results.

Paper, Table 8, PDF p. 21.

# Backup: fewer KV heads work well in the tested small models

| Configuration | Query heads | KV heads | Head dim. | PPL @4K |
|---|---|---|---|---|
| SWA | 12 | 2 | 128 | 11.11 |
| SWA | 6 | 1 | 256 | 11.09 |
| SWA | 12 | 1 | 128 | **10.89** |
| SWA | 12 | 4 | 128 | 11.11 |
| Samba | 12 | 2 | 128 | 10.09 |
| Samba | 6 | 1 | 256 | **9.99** |
| Samba | 12 | 1 | 128 | 10.07 |
| Samba | 12 | 4 | 128 | 10.02 |

MLP width is increased to offset reduced KV parameters; about 430M total parameters.

Paper, Table 6, PDF p. 10. Best rows among the tested configurations, not a universal head-count rule.

# Backup: longer training sequences do not always help

| Training length | Batch size | PPL @4K | PPL @16K |
|---|---|---|---|
| 2,048 | 1,024 | 38.12 | 357.32 |
| 4,096 | 512 | **11.16** | **10.61** |
| 8,192 | 256 | 11.26 | 10.69 |
| 16,384 | 128 | 11.63 | 11.02 |
| 32,768 | 64 | 12.46 | 11.86 |

- Llama-2-SWA 438M; fixed window 2,048; about 2M tokens per step.
- Training length $2w$ is best here; length and batch size change together.
- At length $w$, SWA behaves as full attention throughout training.

Paper, Appendix D, Table 9, PDF p. 22.

# Backup: short convolution improves quality at a speed cost

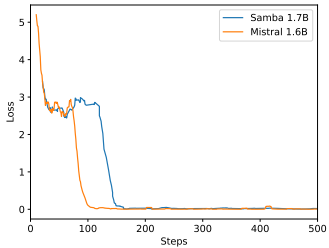
| Model | PPL @4K | PPL @16K | Speed ($10^5$ tokens/s) |
|---|---|---|---|
| Llama-2-SWA | 11.12 | 10.57 | 4.96 |
| + short convolution | 10.83 | 10.31 | 4.69 |
| Sliding GLA | 10.43 | 9.92 | 4.94 |
| + short convolution | 10.39 | 9.87 | 4.44 |
| Sliding RetNet | 10.38 | 9.87 | 4.32 |
| + short convolution | 10.25 | 9.74 | 3.80 |
| Samba (Table 3 reference) | 10.06 | 9.57 | 4.46 |

SC is added to q/k/v; in hybrids, only the linear-attention layers receive it.

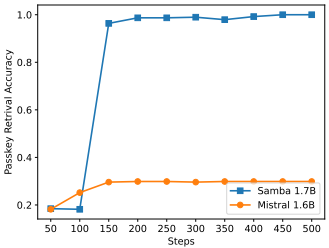
Paper, Appendix D, Table 10, PDF p. 23; Table 3, p. 6. About 430M, measured on 8 A100 GPUs.

# Backup: similar training loss can hide different recall strategies

**4K training loss**




**Long-context retrieval accuracy**




- Both training losses approach zero quickly.
- Samba approaches perfect long-range recall around step 150.
- AdamW, peak LR $10^{-4}$, batch 2,048, 250 warm-up steps.

Paper, Figs. 7–8, PDF pp. 21–22; Appendix C. Fig. 8 caption specifies 256K; protocol also describes seven lengths.

# Backup: selection entropy measures normalized gates

For $h$ heads, sequence length $n$, and generation length $l$:

$$\mathcal{H}_a = -\frac{1}{lh} \sum_{i=1}^{h} \sum_{j=n-l+1}^{n} \sum_{k=1}^{n} A_{ijk} \log A_{ijk}.$$

For Mamba, normalize each channel's gate over *time*:

$$\Delta'_{ij} = \frac{\Delta_{ij}}{\sum_{r=1}^{n} \Delta_{rj}}, \qquad \mathcal{H}_s = -\frac{1}{d_e} \sum_{j=1}^{d_e} \sum_{i=1}^{n} \Delta'_{ij} \log \Delta'_{ij}.$$

- $A_{ijk}$ here denotes attention probabilities, not the Mamba rate parameter.
- $\mathcal{H}_s$ measures temporal gate dispersion, not stored information capacity.

Paper, Appendix E, PDF p. 23.

# Backup: architecture dimensions across scales

| Samba | 421M | 1.3B | 1.7B | 3.8B |
|---|---|---|---|---|
| Total sublayers $N$ | 24 | 36 | 48 | 64 |
| Model width $d_m$ | 1536 | 2304 | 2048 | 2816 |
| MLP width $d_p$ | 4096 | 6144 | 8196 | 9984 |
| Query heads | 12 | 18 | 32 | 11 |
| KV heads | 12 | 18 | 4 | 1 |
| Vocabulary size | 32000 | 32000 | 50304 | 32064 |

- All: $d_e = 2d_m$, $d_s = 16$, $d_r = d_m/16$, convolution width 4.
- All: training length 4,096, attention window 2,048.
- Pure Mamba 1.8B is listed as 64 layers in Table 11, versus 48 in Sec. 2.1.

Paper, Tables 11–12, PDF p. 24; Sec. 2.1.1, p. 3. Detailed baseline configuration follows Table 11.