# Understanding Mamba-3 Through Oscillatory Dynamics

## Paper reading and research discussion

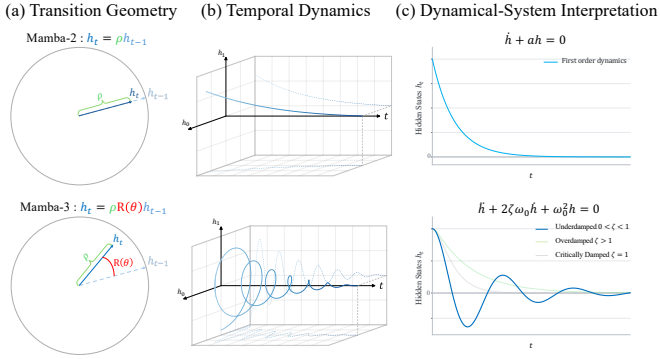
Paper authors: Shutian Liang and Yue Song
College of AI, Tsinghua University

arXiv:2610.13341v1 · Reading: 19 October 2026

Decay controls retention; phase shapes accumulation and access.

Reading presentation based on the paper; not an author presentation. arxiv.org/abs/2610.13341v1

# Rotation changes the direction of recurrent memory




Question: What does learned rotation contribute beyond retaining state magnitude?

Paper, Secs. 1–2, Fig. 1, PDF pp. 1–3. Fixed coefficients and no input in the trajectory illustration.

# A rotary pair separates retention from phase

$$h_t = \bar{A}_t h_{t-1} + \bar{B}_t x_t, \qquad y_t = C_t^\top h_t$$

## Mamba-2: shared decay within a head

$$\bar{A}_t = \alpha_t I$$

Homogeneous propagation preserves direction and scales magnitude.

## Mamba-3: a two-dimensional rotary pair

$$\bar{A}_t = \rho_t R(\theta_t)$$

$\rho_t$: radial retention.
$\theta_t$: input-dependent phase increment.

- SISO and MIMO share the decay–rotation structure.
- Their write/read parameterizations differ; interventions evaluate both.

Paper, Sec. 2.1, Eqs. (2), (4)–(5), PDF pp. 2–3.

# Each frozen interval has an exact continuous flow

$$z_{t+1} = \rho_t R(\theta_t) z_t, \qquad z_t = (H_t, P_t)^\top$$

$$a_t = -\frac{\log \rho_t}{\Delta_t}, \quad \nu_t = \frac{\theta_t}{\Delta_t}, \quad J = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \quad J^2 = -I$$

## Exact matrix exponential

$$A_t^{\mathrm{rot}} = -a_t I + \nu_t J \quad \Longrightarrow \quad e^{A_t^{\mathrm{rot}} \Delta_t} = e^{-a_t \Delta_t} R(\nu_t \Delta_t) = \rho_t R(\theta_t).$$

- Assumptions: $0 < \rho_t < 1$, $\Delta_t > 0$, coefficients fixed within the interval.
- This step analyzes homogeneous propagation, with input injection separated.

Paper, Sec. 3, Eqs. (9)–(13), PDF p. 4; App. C.1, PDF p. 21.

# Nonzero phase gives an underdamped realization

$$\dot{H} = -aH - \nu P, \qquad \dot{P} = \nu H - aP$$

$$\boxed{\ddot{H} + 2a\dot{H} + (a^2 + \nu^2)H = 0}$$

$$\omega_0 = \sqrt{a^2 + \nu^2}, \qquad \zeta = \frac{a}{\sqrt{a^2 + \nu^2}} < 1, \qquad \omega_d = |\nu|$$

- $a$ determines the decay envelope; $|\nu|$ determines oscillation frequency.
- The sign of $\nu$ determines rotation direction.
- For $\nu \neq 0$, recover the second coordinate as $P = -(\dot{H} + aH)/\nu$.

Paper, Sec. 3, Eqs. (14)–(17), PDF p. 4. Equivalence holds within a frozen interval and requires $\nu \neq 0$.

# A shared second-order structure isolates damping regime

$$A_t(\kappa_t) = \begin{bmatrix} 0 & 1 \\ -(a_t^2 - \kappa_t) & -2a_t \end{bmatrix}, \qquad \lambda_\pm = -a_t \pm \sqrt{\kappa_t}$$

| Regime | Constraint | Eigenvalues |
|---|---|---|
| Underdamped | $\kappa_t = -\nu_t^2 < 0$ | Complex conjugate |
| Critical | $\kappa_t = 0$ | Repeated real |
| Overdamped | $0 < \kappa_t < a_t^2$ | Distinct negative real |

- All use $\bar{A}_t = \exp(\Delta_t A_t)$ and coupled two-dimensional states.
- State dimension, input/readout and input-update rule are held unchanged.

Paper, Sec. 4.1, Eqs. (18)–(21), PDF p. 5; App. D.1, PDF p. 24.

# Underdamping preserves relative two-dimensional geometry

**Theory setting:** homogeneous dynamics, fixed coefficients and fixed $\omega_0$.

| Regime | Dominant decay rate | Long-run geometry |
|---|---|---|
| Underdamped | $a < \omega_0$ | Two equally decaying modes |
| Critical | $\omega_0$ | Rescaled operator tends to rank one |
| Overdamped | $a - \sqrt{a^2 - \omega_0^2}$ | Slow mode dominates |

$$z_n = e^{-an\Delta} R(n\nu\Delta) z_0$$

In rotation coordinates: a common scale factor preserves relative angles.

Critical damping has the fastest asymptotic exponential attenuation at fixed $\omega_0$.

Paper, Sec. 4.2, Eqs. (22)–(27), PDF p. 5; App. C.2, PDF pp. 22–23. Overdamped statement assumes a nonzero slow-mode component.

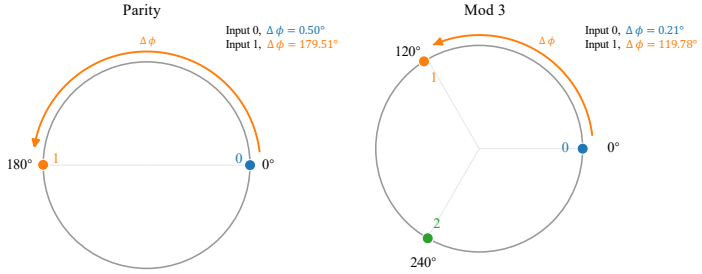
# Input-dependent underdamping wins on state tracking

| Model | Parity | Mod 3 | Arithmetic* |
|---|---|---|---|
| Mamba-2 | 3.91 | 0.04 | 36.76 |
| Mamba-3, no rotation | 3.03 | 0.78 | 48.95 |
| Mamba-3, static rotation | 4.05 | 0.85 | 25.49 |
| Overdamped | 3.27 | 0.58 | 81.60 |
| Critically damped | 3.22 | 0.32 | 1.97 |
| **Underdamped** | **100.00** | **95.26** | **95.68** |

- Training curriculum: maximum length 40 $\rightarrow$ 160; test length **224**.
- Shared $d_{\text{model}} = d_{\text{state}} = 64$; validation selects among eight learning rates.
- Static rotation performs poorly: phase must respond to the incoming symbol.

Paper, Table 2, PDF p. 6; App. D.1, PDF p. 24. Scaled Accuracy (%) ↑; v1 does not spell out its scaling formula.
*Without brackets.

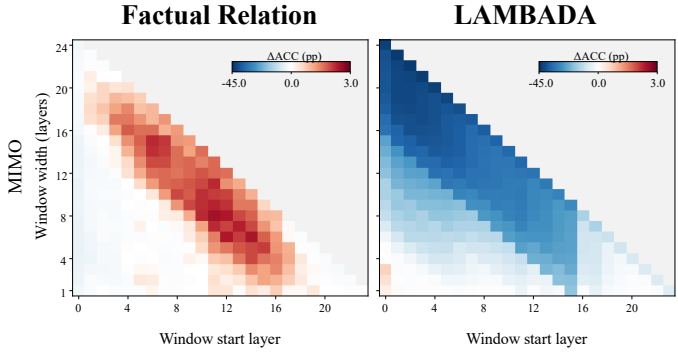
# Selected state pairs learn conditional modular rotations




- Parity: input 0 $\rightarrow 0.50^\circ$; input 1 $\rightarrow 179.51^\circ$.
- Mod 3: input 0 $\rightarrow 0.21^\circ$; input 1 $\rightarrow 119.78^\circ$.
- These selected pairs resemble an ideal count-of-ones phase code: $360^\circ/m$ per 1.

Paper, App. E.1, Fig. 5, PDF p. 27. Selected state pairs, not a claim about every channel.

# Context-dependent prediction is more phase-sensitive




- Remove phase in a contiguous layer window; color is accuracy change in pp.
- Modest effects on Relations; broad losses on LAMBADA. MIMO panels shown; SISO has the same qualitative pattern.

Paper, Sec. 5.1, Fig. 2 (MIMO panels), PDF pp. 6–7. Vector crop; original axes retained and original identical colorbars repeated. Official 1.5B checkpoints.

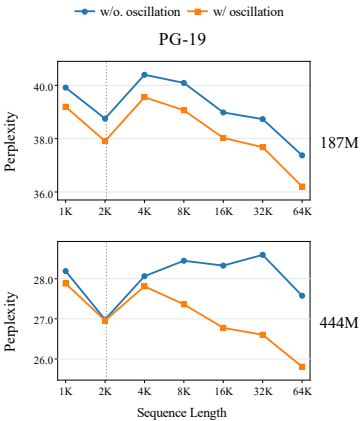
# From-scratch training tests what rotation helps models learn

| Configuration | 187M | 444M |
|---|---|---|
| Layers / model width | 12 / 768 | 24 / 1024 |
| State dimension / head dimension | 128 / 64 | 128 / 64 |
| MIMO rank / chunk size | 4 / 16 | 4 / 16 |
| Rotary channel fraction | 0 versus 0.5 | 0 versus 0.5 |
| Peak learning rate | $3 \times 10^{-3}$ | $1.5 \times 10^{-3}$ |

- Each model: **100B FineWeb-Edu tokens**, Llama-3.1 tokenizer, 2048-token training context.
- AdamW, cosine schedule, global batch 256; BF16 computation with FP32 parameters/states.
- Evaluate PG-19, LongBench-E, standard tasks and four additional corpora.

Paper, Sec. 5.2, PDF p. 7; App. D.3, Table 10, PDF pp. 24–25.

# Oscillation lowers PG-19 perplexity at both scales




- Blue: no oscillation. Orange: oscillation.
- Train at 2K; evaluate up to 64K.
- Reset state per window; score the **last 100 tokens**.
- Lower PPL at both scales; larger long-context gap at 444M.
- No-oscillation models remain stable.

Compare variants at the same length: eligible books and targets vary with window length.

Paper, Fig. 3, PDF p. 7; Sec. 5.3, PDF p. 9; App. D.3, PDF pp. 24–25. PG-19 lengths use $K = 1000$.

# LongBench-E gains appear at 444M, with task trade-offs

| Scale | No oscillation | Oscillation | Difference |
|---|---|---|---|
| 187M | 9.76 | 9.53 | −0.23 |
| 444M | 10.61 | **13.22** | +2.61 |

- Scores average 13 tasks; this is not a single accuracy metric.
- At 444M, major gains: SAMSum +9.25, code completion +9.00 points.
- Passage Count −2.39 and Passage Retrieval −1.24 points.

## Interpretation

Improved language-model perplexity at both scales does not imply uniform downstream gains.

Paper, Sec. 5.2, Table 3, PDF p. 7. Differences calculated from reported scores. Full task names: Table 11, PDF p. 27.

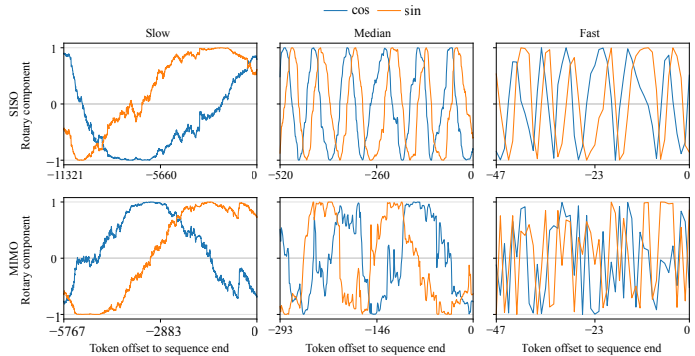
# Standard benchmarks show modest aggregate changes

| Scale | Dynamics | FW-Edu PPL | LAMBADA PPL | Avg. acc. |
|---|---|---|---|---|
| 187M | No oscillation | 16.32 | 38.04 | 43.63 |
| 187M | Oscillation | **16.18** | **32.97** | 43.43 |
| 444M | No oscillation | 12.68 | 18.38 | 50.18 |
| 444M | Oscillation | **12.60** | **17.19** | **51.08** |

- PPL improves at both scales; lower is better.
- Average accuracy changes by $-0.20$ pp at 187M and $+0.90$ pp at 444M.
- The seven-task mean mixes the paper's task-specific `acc` / `acc_n` metrics.

Paper, Table 4, PDF p. 8. Seven tasks: LAMBADA, HellaSwag, PIQA, ARC-E, ARC-C, WinoGrande and OpenBookQA.

# Learned phase varies over multiple timescales




Cosine/sine of accumulated phase; each panel uses its own token-offset range.
Selected near the 10th, 50th and 99th phase-travel percentiles among channels completing a cycle.

Paper, Fig. 4, PDF p. 8; App. D.4, PDF p. 25. PG-19 validation, first 16,000 tokens.

# Interventions perturb distinct parts of a learned computation

- **Phase shuffle:** reassign increments among rotary pairs within each layer, head and token, before accumulation.
- **Phase removal:** set raw angle projections to zero.
- **Phase reversal:** negate every phase increment.
- **Decay permutation:** cyclically reassign decay coefficients $A$ across heads within a layer, keeping the receiving head's $\Delta$.

$$\widetilde{(A\Delta)}_{\ell,h,t} = A_{\ell,\pi_\ell(h),t}\Delta_{\ell,h,t}$$

Protocol: official 1.5B checkpoints; all layers and the whole window; matched windows and scored tokens within each context length.

Paper, Sec. 5.3, Table 5, PDF p. 8; App. D.4, PDF pp. 25–26. Phase interventions preserve radial decay. Perturbation strengths are not matched.

# SISO becomes increasingly sensitive to phase disruption

| Intervention | 2K | 16K | 64K |
|---|---|---|---|
| Native PPL | 21.09 | 21.41 | 20.60 |
| Decay permutation | 5.43× | 5.91× | 6.08× |
| Phase shuffle | 2.38× | 4.49× | 7.25× |
| Phase removal | 3.22× | 13.34× | 41.33× |
| Phase reversal | 7.22× | 27.93× | 74.10× |

Ratios are $\mathrm{PPL}_{\mathrm{intervened}}/\mathrm{PPL}_{\mathrm{native}}$; larger means more damage.

At 64K, phase removal gives PPL **851.49**; reversal gives **1526.52**.

Paper, Table 5, PDF p. 8. Frozen official 1.5B SISO checkpoint; this is an inference-time intervention.

# MIMO has a different pattern of phase sensitivity

| Intervention | 2K | 16K | 64K |
|---|---|---|---|
| Native PPL | 21.12 | 21.47 | 20.52 |
| Decay permutation | 1.22× | 1.21× | 1.21× |
| Phase shuffle | 3.05× | 3.28× | 3.20× |
| Phase removal | 9.34× | 14.92× | 15.08× |
| Phase reversal | 3.52× | 3.67× | 3.60× |

- Removal is most damaging; reversal and shuffle show weaker length dependence.
- Phase interventions are more damaging than this decay-permutation control.

Paper, Table 5, PDF p. 8. Ratios relative to native PPL at each length; official 1.5B MIMO checkpoint.

# Removing phase increases alignment and state magnitude

$$S_t = P_t + U_t, \qquad \|S_t\|_F^2 = \|P_t\|_F^2 + \|U_t\|_F^2 + 2\langle P_t, U_t\rangle_F$$

| Model | Length | Cosine: native → remove | RMS: native → remove |
|---|---|---|---|
| SISO | 16K | 0.293 → 0.429 | 7.635 → 8.899 |
| SISO | 64K | 0.294 → 0.474 | 7.621 → 9.437 |
| MIMO | 16K | 0.241 → 0.488 | 10.073 → 12.022 |
| MIMO | 64K | 0.242 → 0.499 | 10.078 → 12.047 |

- Larger positive alignment favors constructive accumulation.
- Supports the authors' explanation that rotation regulates how updates combine.

Paper, Table 6, PDF p. 9; App. D.4, PDF p. 26. Norm expansion is an explanatory identity. Statistics exclude the first 2,000 tokens; RMS is not dimension-normalized.

# Phase can control accumulation without changing retention

**Fixed-coefficient toy system with constant input:**

$$z_t = \rho e^{i\theta} z_{t-1} + u, \qquad 0 \le \rho < 1, \quad z_0 = 0$$

$$\boxed{\|z_\infty\| = \frac{\|u\|}{|1 - \rho e^{i\theta}|}} \qquad |1 - \rho e^{i\theta}|^2 = (1-\rho)^2 + 2\rho(1 - \cos\theta)$$

## Illustrative calculation, not an experimental result

For $\rho = 0.99$ and $\|u\| = 1$:

$$\theta = 0: \quad \|z_\infty\| = 100 \qquad\qquad \theta = \pi/2: \quad \|z_\infty\| \approx 0.711$$

Rotation changes accumulation while preserving the radial retention factor.

Paper, App. C.3, Eqs. (115)–(120), PDF p. 23. Denominator expansion and numerical example added for explanation.

# Writing and readout need compatible phase coordinates

**64K windows:** native history; intervene only during the last 100 tokens.

| Tail intervention | SISO PPL | MIMO PPL |
|---|---|---|
| Native | 20.60 | 20.52 |
| Remove rotation from $B$ | 186.28 (9.04×) | 100.41 (4.89×) |
| Remove rotation from $C$ | 250.90 (12.18×) | 117.68 (5.74×) |
| Remove rotation from both | 42.79 (2.08×) | 59.46 (2.90×) |

- Write/read vectors remain present; only their rotary transformations change.
- Joint removal partially restores performance, supporting phase compatibility.
- Earlier history still uses native coordinates, which may explain a residual gap.

Paper, Table 7, PDF p. 9; App. D.4, PDF p. 26. Native state is carried into the intervention region without resetting.

# Phase shapes computation, accumulation and access

1. **Compute:** input-dependent underdamping supports state tracking beyond two-coordinate coupling.
2. **Combine:** rotation changes alignment between retained history and incoming writes; both scales improve long-context PPL.
3. **Access:** disrupting phase or write/read compatibility damages pretrained prediction.

## How to read the evidence

Inference-time removal tests learned dependence. From-scratch training tests architectural benefit. Non-oscillatory models can still extrapolate stably.

Liang and Song, Understanding Mamba-3 Through Oscillatory Dynamics, arXiv:2610.13341v1. Secs. 4–6, PDF pp. 5–9.

# Backup: phase removal is not critical damping

## Remove phase in rotary coordinates

$$\nu = 0 \quad \Rightarrow \quad A^{\mathrm{rot}} = -aI$$

$$e^{A^{\mathrm{rot}}\Delta} = e^{-a\Delta} I$$

Two uncoupled equal-rate decays.

## Critical second-order construction

$$A(0) = \begin{bmatrix} 0 & 1 \\ -a^2 & -2a \end{bmatrix}$$

$$e^{A(0)\Delta} = e^{-a\Delta}(I + \Delta N)$$

$N \neq 0$ and $N^2 = 0$: a nontrivial Jordan block.

- The coordinate map $P = -(\dot{H} + aH)/\nu$ becomes singular at $\nu = 0$.
- Fixed-frequency attenuation theory also differs from the learned spectral-family comparison.

Reading analysis from Sec. 3, Eqs. (14)–(17), PDF p. 4, and Sec. 4.1–4.2, PDF p. 5.

# Backup: LongBench-E gains are concentrated at 444M

**Largest gains (score points)**

| Task | Change |
|---|---|
| SAMSum | +9.25 |
| Long Code Completion | +9.00 |
| TREC | +7.33 |
| RepoBench-P | +7.12 |
| TriviaQA | +2.27 |

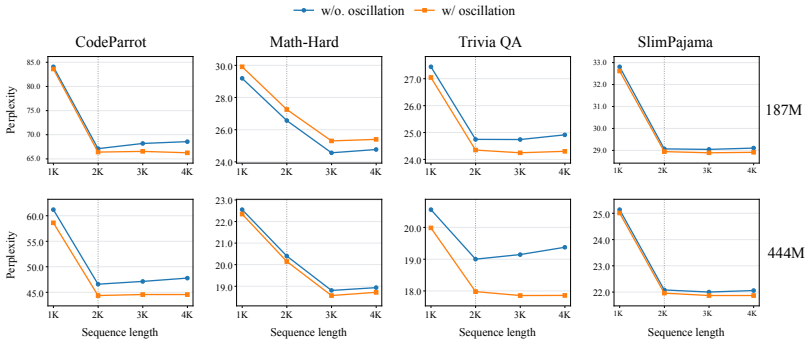
**Regressions (score points)**

| Task | Change |
|---|---|
| Passage Count | −2.39 |
| Passage Retrieval | −1.24 |
| MultiNews | −0.29 |
| Qasper | −0.22 |

Overall: 10.61 → 13.22; synthetic state-tracking success does not imply gains on every long-context task.

Paper, Table 3, PDF p. 7. Selected rows; changes calculated as oscillatory minus non-oscillatory scores.

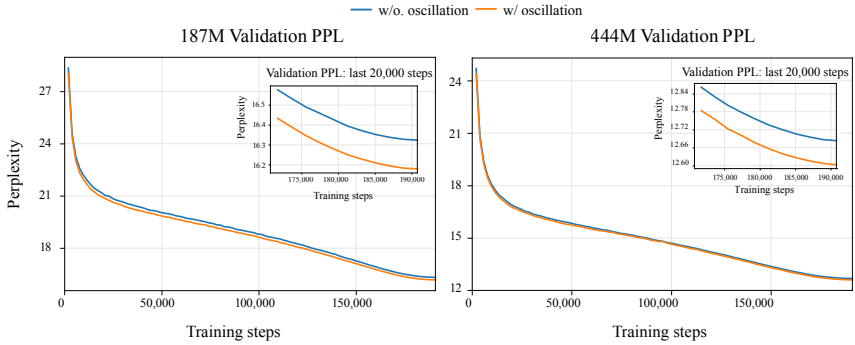
# Backup: additional corpora show scale-dependent benefits




187M loses on MATH-Hard; 444M improves on all four datasets.
The x-axis indexes consecutive 1024-token bins of 4096-token streams; state is carried across bins.

Paper, App. E.2, Fig. 6, PDF pp. 27–28; evaluation protocol: App. D.3, PDF p. 25.

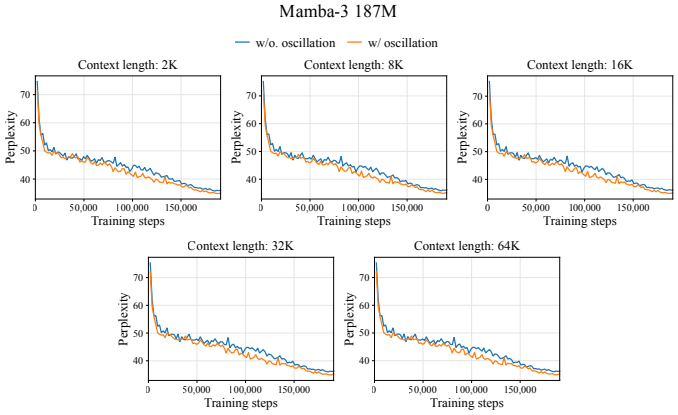
# Backup: validation PPL separates late in pretraining




Both scales end with lower validation PPL for the oscillatory model. Insets show the final 20,000 training steps.

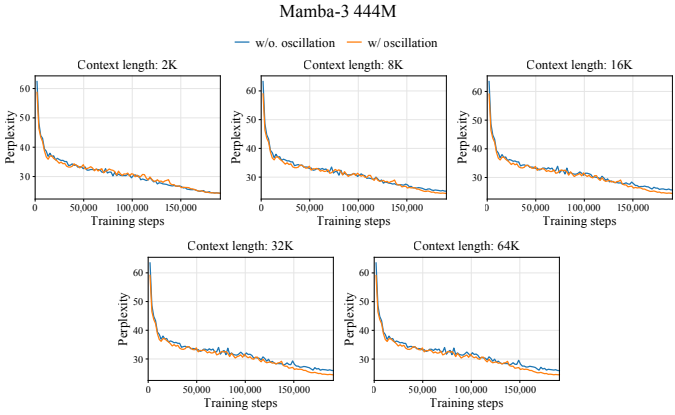
Paper, App. E.3, Fig. 7, PDF p. 28. Horizontal axis: training steps; vertical axis: validation perplexity.

# Backup: 187M long-context gains develop during training




Paper, App. E.3, Fig. 8, PDF p. 29. PG-19 PPL at training checkpoints; panels: 2K, 8K, 16K, 32K and 64K.

# Backup: 444M separates more at long evaluation lengths




Paper, App. E.3, Fig. 9, PDF p. 29. Blue: no oscillation; orange: oscillation. Training uses a 2K context.

# Backup: frozen spectra connect recurrent architectures

$$S_t = T_t S_{t-1} + U_t, \quad H = I - \beta k k^\top, \quad D = \mathrm{Diag}(\boldsymbol{\alpha})$$

| Model | Frozen transition | Modal structure |
|---|---|---|
| Linear Attention | $I$ | Preserved modes |
| DeltaNet | $H$ | Decay plus preserved modes |
| Gated DeltaNet | $\alpha H$ | Two non-oscillatory decay rates |
| KDA | $HD$ | Positive real spectrum |
| Complex KDA | $HD$, signed gates | Can admit oscillatory modes |
| Momentum DeltaNet | Augmented $(S, M)$ | Can admit oscillatory modes |

Unit-norm keys; ordinary DeltaNet/Gated DeltaNet/KDA use $0 < \beta < 1$ and positive decay gates. These are frozen-interval statements.

Paper, App. A, Table 8, PDF pp. 15–20. DeltaNet/Gated DeltaNet pairings share one key-direction mode; they are not independent state pairs.

# Backup: signed gates and momentum allow complex modes

**Complex KDA construction**

$$T_{\alpha,\theta} = R(\theta)\,\mathrm{Diag}(-\alpha, 1), \quad -1 < \alpha < 0$$

$$\mathcal{D} = (1-\alpha)^2 \cos^2\theta + 4\alpha < 0 \quad \Rightarrow \quad |\lambda_\pm| = \sqrt{-\alpha} < 1$$

**Momentum DeltaNet: key-direction state–momentum block**

$$B = \begin{bmatrix} \alpha(1-\beta\eta) & -\beta\mu \\ \alpha\eta & \mu \end{bmatrix}, \quad \tau = \alpha(1-\beta\eta) + \mu$$

$$\tau^2 < 4\alpha\mu, \quad 0 < \alpha, \mu < 1 \quad \Rightarrow \quad |\lambda_\pm| = \sqrt{\alpha\mu} < 1$$

These conditions admit an underdamped realization in suitable real modal coordinates.

Paper, App. A.6–A.7, PDF pp. 19–20; App. C.4, PDF pp. 23–24. Possibility conditions, not guarantees for every update.

# Backup: what the state measurements average

$$c_t = \frac{\langle P_t, U_t \rangle_F}{\|P_t\|_F \|U_t\|_F + 10^{-12}}, \qquad P_t = \alpha_t S_{t-1}$$

- PG-19 validation books with at least 64,000 tokens; one uninterrupted trajectory per condition, starting from zero state.
- Cosine: average over positions $[2000, L)$, median over $24 \times 64$ layer–head pairs, then equal-weight mean across books.
- RMS: average squared Frobenius norms over positions, median over layer–head pairs, square root, then mean across books.
- All state coordinates are included; there is no dimension normalization.

Mechanistic boundary: altered alignment and norms support an accumulation account, while encoding and readout also change under phase removal.

Paper, App. D.4, PDF p. 26; Table 6, PDF p. 9. $P_t$ here is propagated matrix state, not the scalar coordinate in Sec. 3.