UNDERSTANDING MAMBA-3 THROUGH OSCILLATORY DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Mamba-3 introduces input-dependent rotations into recurrent state updates, but their contribution beyond state retention remains unclear. We interpret these rotations as **underdamped second-order dynamics**, providing a framework for investigating the role of oscillation in sequence modeling. Controlled comparisons show that underdamped dynamics outperform non-oscillatory damping regimes on state-tracking tasks, while language-model pretraining demonstrates benefits for long-context prediction and length generalization. Interventions in pretrained models further establish the importance of learned phase, and state analysis suggests that rotation regulates how retained information combines with incoming updates. Taken together, these findings identify underdamped dynamics as a useful mechanism for recurrent computation and suggest that oscillation supports long-context processing by regulating how information accumulates.

1 INTRODUCTION

State-space models (SSMs) process sequences through recurrent updates to a hidden state (Gu et al., 2021). Their dynamical structure enables linear-time sequence processing and constant-memory autoregressive inference, while structured algorithms support parallel training. These properties make SSMs an efficient alternative to attention-based models (Vaswani et al., 2017). In selective SSMs such as Mamba-1 (Gu & Dao, 2023) and Mamba-2 (Dao & Gu, 2024), input-dependent updates determine what information enters and leaves the state, while multiplicative decay controls how much of the existing state is retained.

Mamba-3 (Lahoti et al., 2026) adds input-dependent rotation to this state evolution through complex-valued transitions. Unlike conventional rotary positional embeddings in Transformers (Su et al., 2024), whose angles follow a prescribed positional schedule, Mamba-3 learns phase increments from the input. As illustrated in Fig. 1, these rotations change the direction of the retained state, turning pure decay into spiraling evolution under fixed coefficients. This geometric distinction raises a natural computational question: *what do learned rotations contribute beyond state retention, particularly over long contexts?*

We investigate this question by interpreting Mamba-3's decay–rotation transition as an **underdamped second-order system**. In this representation, decay sets damping envelopes, while rotation produces oscillatory phase evolution. The connection provides a unified framework for comparing underdamped, critically damped, and overdamped dynamics. Our analysis characterizes how these regimes differ in mode persistence and state geometry, and controlled state-tracking experiments show a clear advantage of underdamped dynamics. Since all three variants retain coupled second-order states, the comparisons highlight the computational contribution of the oscillatory regime.

Our language-modeling experiments examine whether this contribution extends to the use of contextual information. In pretrained Mamba-3 models, removing phase rotation affects context-dependent prediction on LAMBADA (Paperno et al., 2016) much more strongly than factual relation prediction (Hernandez et al., 2024). We further train oscillatory and non-oscillatory models from scratch to assess how rotation affects learned long-context capabilities. Experimental results show that oscillatory models achieve lower PG-19 perplexity (Rae et al., 2019) beyond the training context length at both evaluated scales, with downstream gains on LongBench-E (Bai et al., 2024) at the larger scale.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

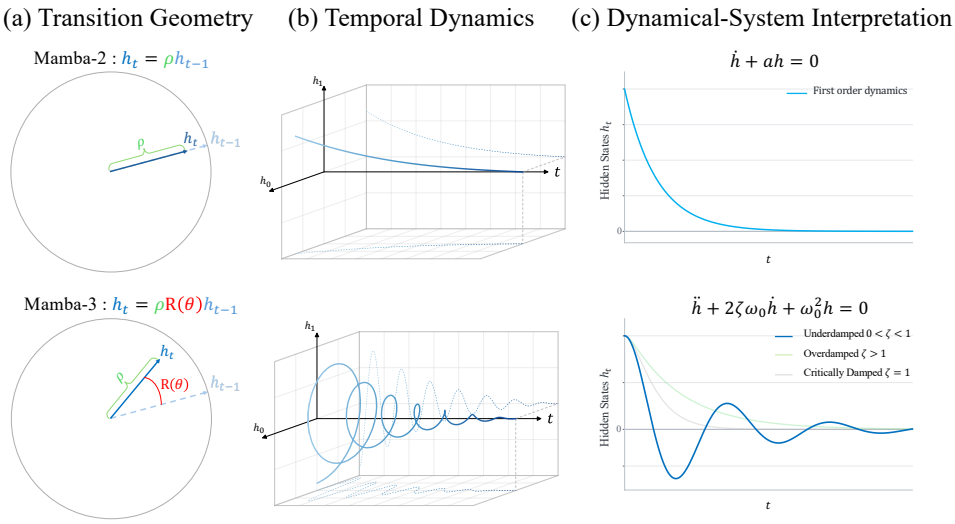

Figure 1: **A geometric perspective on Mamba-3's recurrent dynamics.** (a) Mamba-2 scales the retained state, while Mamba-3 additionally rotates it. (b) With fixed coefficients and no input, decay–rotation produces a spiraling trajectory. (c) This geometry motivates our interpretation of Mamba-3 as an underdamped second-order system with oscillatory dynamics.

To understand how oscillation contributes to these gains, we examine how learned phase shapes recurrent computation through interventions in pretrained Mamba-3 models. Removing, reversing, or shuffling phase increments substantially degrades long-context prediction. Phase removal also increases alignment between retained history and incoming updates, together with larger state norms, suggesting that rotation moderates constructive accumulation. Interventions on writing and readout further reveal the importance of compatible phase coordinates for accessing retained information.

Our main findings are:

- **Underdamped dynamics provide a useful interpretation of Mamba-3.** The second-order representation connects its learned rotations to oscillation and enables damping-regime comparisons that demonstrate the benefit of underdamping for state tracking.
- **Oscillation benefits long-context language modeling.** Controlled pretraining shows improved length generalization at both evaluated scales (187M and 444M), with downstream long-context gains emerging at the larger scale.
- **Learned phase regulates recurrent state accumulation.** Phase interventions establish its importance for long-context prediction, while state analysis suggests that rotation moderates constructive accumulation between retained history and incoming updates.

Together, these findings suggest that effective recurrent memory in SSMs depends on both how much information is retained and how that information is transformed as new inputs arrive. More broadly, this dynamical perspective motivates oscillatory dynamics as a design dimension for sequence architectures, with extensions to linear-attention and other recurrent models (Appendix A).

2 Background

2.1 Selective State-Space Models

State-space models process sequences by maintaining a recurrent hidden state. In continuous time, their dynamics take the form

$$\dot{h}_t = A_t h_t + B_t x_t, \qquad y_t = C_t^\top h_t, \tag{1}$$

where $x(t), y(t) \in \mathbb{R}^d$ and $h(t) \in \mathbb{R}^h$. The transition matrix $A(t) \in \mathbb{R}^{h \times h}$ governs the evolution of the existing state, while $B(t), C(t) \in \mathbb{R}^{h \times d}$ control input injection and output readout. Discretizing over intervals of length Δ_t yields a recurrence of the form

$$h_t = \bar{A}_t h_{t-1} + \bar{B}_t x_t, \qquad y_t = C_t^\top h_t. \tag{2}$$

Selective SSMs make components of this recurrence input-dependent, allowing each token to influence how information is retained, written, and read. We focus on the transition $\bar{A}_t$, whose structure distinguishes the state dynamics of Mamba-1, Mamba-2, and Mamba-3.

Mamba-1: selective diagonal decay. Mamba-1 makes Δ_t, B_t, and C_t input-dependent while retaining a learned, time-invariant diagonal transition matrix A. Its discrete transition is

$$\bar{A}_t = \exp(\Delta_t A), \tag{3}$$

With negative diagonal entries in A, each state coordinate decays at its own rate, and the input-dependent step size Δ_t modulates these rates across tokens.

Mamba-2: shared decay within each head. Mamba-2 restricts the continuous-time transition within each head to $A = aI$, where $a < 0$. The discrete update therefore becomes

$$\bar{A}_t = \alpha_t I, \qquad \alpha_t = \exp(\Delta_t a), \qquad h_t = \alpha_t h_{t-1} + \Delta_t B_t x_t. \tag{4}$$

The scalar α_t controls retention uniformly across state coordinates within a head. Consequently, homogeneous propagation changes the magnitude of the state while preserving its direction, while token-dependent writing and readout remain controlled by B_t and C_t.

Mamba-3: decay with input-dependent rotation. Mamba-3 introduces a phase degree of freedom that allows the propagated state to change direction as well as decay. For each two-dimensional rotary state pair, the transition takes the form

$$\bar{A}_t = \rho_t R(\theta_t), \qquad R(\theta_t) = \begin{bmatrix} \cos\theta_t & -\sin\theta_t \\ \sin\theta_t & \cos\theta_t \end{bmatrix}, \tag{5}$$

where ρ_t controls radial retention and θ_t is an input-dependent rotation angle, implemented through data-dependent rotary embeddings. Mamba-3's single-input, single-output (SISO) and multi-input, multi-output (MIMO) variants share this decay–rotation structure but differ in their input and readout parameterizations. Our analysis mainly focuses on the dynamical role of the learned phase.

2.2 Damped Second-Order Dynamics

A damped second-order system evolves according to

$$\ddot{x}(t) + 2\zeta\omega_0\dot{x}(t) + \omega_0^2 x(t) = u(t), \tag{6}$$

where $\omega_0 > 0$ is the natural frequency, $\zeta > 0$ is the damping ratio, and $u(t)$ is an external forcing term. Introducing the velocity $v(t) = \dot{x}(t)$ gives the first-order state-space representation

$$\frac{d}{dt}\begin{bmatrix} x(t) \\ v(t) \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 1 \\ -\omega_0^2 & -2\zeta\omega_0 \end{bmatrix}}_{A_{\mathrm{SO}}} \begin{bmatrix} x(t) \\ v(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t). \tag{7}$$

The eigenvalues of A_{SO} determine the system's intrinsic response:

$$\lambda_\pm = -\zeta\omega_0 \pm \omega_0\sqrt{\zeta^2 - 1}. \tag{8}$$

Table 1 summarizes three regimes: underdamped systems have complex-conjugate eigenvalues and exhibit decaying oscillations, whereas critically damped and overdamped systems have real eigenvalues and non-oscillatory intrinsic dynamics. This distinction allows us to examine the role of oscillation within a common second-order framework.

Table 1: Damping regimes of a second-order dynamical system.

Regime	Eigenvalues	Dynamics
Underdamped ($0 < \zeta < 1$)	Complex conjugate	Oscillatory
Critically damped ($\zeta = 1$)	Repeated real	Non-oscillatory
Overdamped ($\zeta > 1$)	Distinct real	Non-oscillatory

3 AN OSCILLATORY INTERPRETATION OF MAMBA-3

We now derive a second-order representation of Mamba-3's decay–rotation transition and identify its underlying damping regime. The analysis considers homogeneous state propagation, separating the intrinsic transition dynamics from input injection.

From discrete rotation to continuous dynamics. Consider a rotary state pair $z_t = (H_t, P_t)^\top$ evolving as

$$z_{t+1} = \rho_t R(\theta_t) z_t, \qquad 0 < \rho_t < 1, \tag{9}$$

where $R(\theta_t)$ is the two-dimensional rotation matrix defined in Eq. 5. For an interval of length $\Delta_t > 0$, define the decay rate and angular frequency as

$$a_t = -\frac{\log \rho_t}{\Delta_t}, \qquad \nu_t = \frac{\theta_t}{\Delta_t}. \tag{10}$$

The corresponding continuous-time generator is

$$A_t^{\mathrm{rot}} = -a_t I + \nu_t J, \qquad J = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \tag{11}$$

Since $J^2 = -I$, its exponential satisfies

$$e^{\phi J} = \cos(\phi) I + \sin(\phi) J = R(\phi), \tag{12}$$

as derived in Appendix C.1. Because I and J commute,

$$\begin{aligned} e^{A_t^{\mathrm{rot}} \Delta_t} &= e^{-a_t \Delta_t} e^{\nu_t \Delta_t J} \\ &= e^{-a_t \Delta_t} \left[\cos(\nu_t \Delta_t) I + \sin(\nu_t \Delta_t) J\right] \\ &= e^{-a_t \Delta_t} \begin{bmatrix} \cos\theta_t & -\sin\theta_t \\ \sin\theta_t & \cos\theta_t \end{bmatrix} = \rho_t R(\theta_t). \end{aligned} \tag{13}$$

Thus, each discrete decay–rotation update is the exact flow of $\dot{z} = A_t^{\mathrm{rot}} z$, with coefficients held fixed over that interval.

An underdamped second-order representation. Within one interval, suppress the index t and write the continuous dynamics componentwise:

$$\dot{H} = -aH - \nu P, \qquad \dot{P} = \nu H - aP. \tag{14}$$

Differentiating the first equation and substituting the second yields

$$\ddot{H} = -a\dot{H} - \nu(\nu H - aP) = -2a\dot{H} - (a^2 + \nu^2) H, \tag{15}$$

where we used $\nu P = -\dot{H} - aH$. Hence,

$$\ddot{H} + 2a\dot{H} + (a^2 + \nu^2) H = 0. \tag{16}$$

For $\nu \neq 0$, the auxiliary coordinate is recovered as $P = -(\dot{H} + aH)/\nu$, so the two representations are equivalent within the interval.

Comparing with the canonical damped second-order equation gives

$$\omega_0 = \sqrt{a^2 + \nu^2}, \qquad \zeta = \frac{a}{\sqrt{a^2 + \nu^2}}, \qquad \omega_d = \omega_0 \sqrt{1 - \zeta^2} = |\nu|. \tag{17}$$

Since $a > 0$, any nonzero ν gives $0 < \zeta < 1$, placing the dynamics in the underdamped regime. The decay rate a determines the damping envelope, while $|\nu|$ sets the oscillation frequency and the sign of ν specifies the rotation direction. Mamba-3's learned phase therefore controls the oscillatory component of state evolution. In Sec. 4.1, we compare this regime with critically damped and overdamped alternatives to assess the contribution of oscillation to recurrent computation.

4 OSCILLATION AND STATE TRACKING

Does the benefit of Mamba-3's rotary transition arise from oscillation or simply from coupling two state coordinates? We investigate this question by comparing oscillatory and non-oscillatory variants with the same second-order state structure. State-tracking tasks provide a controlled setting in which models must maintain and update an internal state as new tokens arrive.

4.1 Controlled Comparisons Across Damping Regimes

To construct these variants, we generalize second-order dynamics in Eq. 16 through a parameter κ_t:

$$A_t(\kappa_t) = \begin{bmatrix} 0 & 1 \\ -(a_t^2 - \kappa_t) & -2a_t \end{bmatrix}, \tag{18}$$

Within each interval, this generator corresponds to

$$\ddot{H} + 2a_t\dot{H} + (a_t^2 - \kappa_t)H = U_t, \tag{19}$$

where U_t denotes an external forcing term. Its eigenvalues are:

$$\lambda_\pm = -a_t \pm \sqrt{\kappa_t}. \tag{20}$$

Thus, κ_t controls the modal structure around the shared spectral center $-a_t$. Setting $\kappa_t = -\nu_t^2 < 0$ recovers the underdamped representation derived for Mamba-3 within each interval, while $\kappa_t = 0$ and $0 < \kappa_t < a_t^2$ yield critically damped and overdamped alternatives, respectively.

All variants retain the coupled two-dimensional state structure. We use the same input-dependent step size and discretize each generator as

$$\bar{A}_t(\kappa_t) = \exp(\Delta_t A_t(\kappa_t)). \tag{21}$$

The state dimension, input and readout, and input-update rule are kept unchanged across variants. This comparison tests the benefit of oscillatory modes beyond second-order state coupling alone.

4.2 Mode Persistence and State Geometry

We next examine how each damping regime affects the persistence and geometry of propagated states. We consider homogeneous dynamics with fixed coefficients, holding the natural frequency $\omega_0 > 0$ constant while varying $a > 0$ and setting $\kappa = a^2 - \omega_0^2$. The continuous-time generator and its propagation operator are

$$A(a) = \begin{bmatrix} 0 & 1 \\ -(a^2 - \kappa) & -2a \end{bmatrix}, \qquad \Phi_a(t) = e^{A(a)t}, \tag{22}$$

Overdamped dynamics: dominance of a single mode. For $a > \omega_0$, we have $\kappa > 0$ and eigenvalues $\lambda_\pm = -a \pm \sqrt{\kappa}$. With a fixed step size $\Delta > 0$, the discrete eigenvalues $\mu_\pm = e^{\lambda_\pm \Delta}$ satisfy

$$\lim_{n\to\infty} |\frac{\mu_-}{\mu_+}|^n = \lim_{n\to\infty} e^{-2\sqrt{\kappa}\, n\Delta} = 0. \tag{23}$$

The fast mode therefore becomes negligible relative to the slow mode as the number of recurrent steps n increases. For initial states with a nonzero slow-mode component, propagation becomes dominated by the slow eigendirection, indicating *asymptotic mode collapse*.

Critically damped dynamics: fastest asymptotic attenuation. At $a = \omega_0$, the eigenvalues coincide at $-\omega_0$. Writing $A = -\omega_0 I + N$, where $N^2 = 0$ and $N \neq 0$, gives

$$\Phi_{\omega_0}(t) = e^{-\omega_0 t}(I + tN). \tag{24}$$

The polynomial factor does not change the asymptotic exponential decay rate, which is ω_0. By comparison, the dominant overdamped mode decays at rate

$$a - \sqrt{a^2 - \omega_0^2} = \frac{\omega_0^2}{a + \sqrt{a^2 - \omega_0^2}} < \omega_0, \tag{25}$$

while the underdamped decay rate is $a < \omega_0$. Critical damping therefore yields the *fastest asymptotic exponential attenuation* at a fixed natural frequency. Its propagation also becomes directionally degenerate, since $e^{\omega_0 t}\Phi_{\omega_0}(t)/t \to N$, a rank-one matrix.

Underdamped dynamics: persistent two-dimensional geometry. For $0 < a < \omega_0$, let $\nu = \sqrt{\omega_0^2 - a^2}$. The eigenvalues are $\lambda_\pm = -a \pm i\nu$, giving

$$|\mu_+| = |\mu_-| = e^{-a\Delta}, \qquad \lim_{n\to\infty} |\frac{\mu_-}{\mu_+}|^n = 1. \tag{26}$$

Neither mode is preferentially attenuated. In the rotation coordinates, propagation takes the form

$$z_n = e^{-an\Delta} R(n\nu\Delta) z_0. \tag{27}$$

Thus, lengths and angles are preserved up to a common radial decay, without progressive directional collapse. Since $a < \omega_0$, attenuation is also slower than under critical damping. Underdamped dynamics therefore combine *preservation of relative state geometry* with *slower asymptotic decay than critical damping*. Full proofs of mode persistence analysis are provided in Appendix C.2.

4.3 State-Tracking Tasks

Experimental Setup. We evaluate Parity, Mod 3, and Arithmetic without Brackets through two comparisons: varying the damping regime within the shared second-order framework, comparing Mamba-3 with fixed and absent rotations. These comparisons assess the importance of underdamped and input-dependent phase. We also include Mamba-2 as a baseline.

Table 2: Scaled Accuracy (%) of state-tracking.

Model	Parity	Mod 3	Arith. w/o Brackets
Mamba-2	3.91	0.04	36.76
Mamba-3 w/o rotation	3.03	0.78	48.95
Mamba-3 w/ Static rotation	4.05	0.85	25.49
Overdamped	3.27	0.58	81.60
Critically Damped	3.22	0.32	1.97
Underdamped (Mamba-3)	**100.0**	**95.26**	**95.68**

Results. Table 2 shows a consistent advantage of underdamped dynamics over critically damped and overdamped alternatives. The gap is particularly pronounced on Parity and Mod 3, whose cyclic structure naturally accommodates rotational updates. The advantage extends to Arithmetic without Brackets, where the underdamped variant achieves 95.68% scaled accuracy, compared with 81.60% for the overdamped variant. Since these variants share the same second-order state structure, the results highlight the importance of the *underdamped regime* for state tracking. The rotation comparisons further show that *data dependence matters*: Mamba-3 substantially outperforms variants with fixed or removed rotations across all three tasks. Thus, effective state tracking benefits from both oscillatory dynamics and the ability to adapt phase to incoming tokens.

> **Main Findings: Oscillation supports effective state tracking when its phase adapts to incoming information.** Underdamped dynamics outperform critically damped and overdamped alternatives, while data-dependent rotations outperform fixed rotations.

5 Oscillation in Language Modeling

We next examine which language tasks depend on oscillatory dynamics, whether oscillation improves long-context performance, and how learned phase shapes recurrent state propagation.

5.1 Oscillation across Language Tasks

Experimental setup. We compare factual relation prediction on Relation (Hernandez et al., 2024) with context-dependent word prediction on LAMBADA (Paperno et al., 2016) to assess how reliance on oscillation varies across tasks. Following Sharma et al. (2024), we set the rotation phase to zero within consecutive layer windows in pretrained 1.5B Mamba-3 SISO and MIMO models. We vary the window position and size and measure accuracy changes relative to the unmodified models.

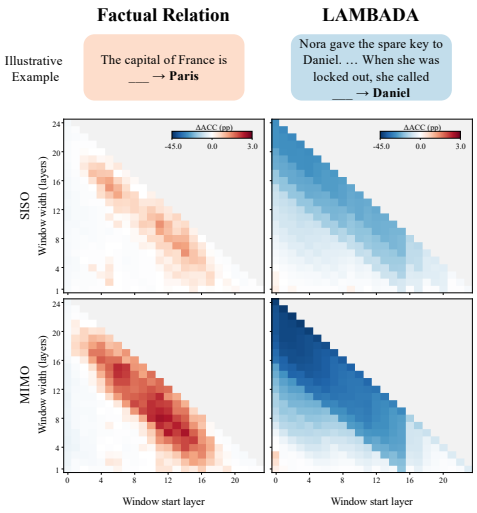

Figure 2: Accuracy changes after phase removal in contiguous layer windows.

Results. Removing phase rotation has a much stronger effect on LAMBADA than on Relation (Fig. 2). Relation accuracy changes only modestly,

with occasional improvements when rotation is removed from middle layers. LAMBADA accuracy, however, drops substantially across many window positions, and removing rotation from more layers generally increases the degradation. This task-dependent sensitivity suggests that learned phase is particularly important for using contextual information, motivating a closer examination of its role in long-context language modeling.

> **Main Findings: Learned oscillations are particularly important for context-dependent prediction.** Removing phase rotation has only a modest effect on factual relation prediction but substantially degrades performance on LAMBADA, which requires integrating information across the preceding context.

5.2 Pretraining and Long-Context Evaluation

Experimental setup. We test whether the importance of phase in pretrained models translates into better long-context performance when models are trained from scratch. We train Mamba-3 variants with/without oscillations at the scale of 187M and 444M, each on 100B tokens of FineWeb-Edu dataset (Penedo et al., 2024) with a 2K context length, following Lahoti et al. (2026). We evaluate length generalization through PG-19 perplexity (Rae et al., 2019) at context lengths up to 64K and downstream tasks on LongBench-E (Bai et al., 2024).

Results. As shown in Fig. 3, oscillatory models achieve lower PG-19 perplexity at both scales, with a more pronounced advantage at longer contexts for the larger model. We find that this benefit is not specific to PG-19: following Grazzi et al. (2025), we further evaluate perplexity on CodeParrot (Tunstall et al., 2022), MATH-Hard (Hendrycks et al., 2021), TriviaQA (Joshi et al., 2017), and SlimPajama (Soboleva et al., 2023), finding lower perplexity beyond the training context on most datasets (Appendix E.2).

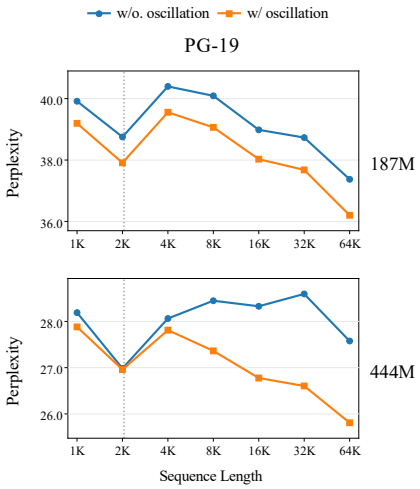

Figure 3: PG-19 perplexity across context lengths for models with/without oscillation. The dotted line marks the 2K training context length.

On LongBench-E, the variants perform comparably at 187M, scoring 9.53 with oscillation and 9.76 without. At 444M, oscillation increases the average score from 10.61 to **13.22** (Table 3). These results distinguish the improvements in language-modeling perplexity observed at both scales from the downstream gains observed at 444M. Standard benchmarks provide a complementary check

Table 3: LongBench-E results for models with and without oscillatory dynamics. We compare the non-oscillatory variant and the default oscillatory configuration at two model scales.

Params	Dynamics	Synthetic		Summary		Single-doc QA		Multi-doc QA		Few-shot Learning			Coding		Avg.
		PC	PR	GR	MN	MQA	QA	2WM	HQA	SS	TR	TQA	LCC	RB	
187M	w/o oscillation	**5.96**	6.07	7.75	9.17	9.62	**4.37**	**6.71**	**5.24**	7.87	**12.67**	10.63	**20.25**	**20.57**	**9.76**
	w/ oscillation	5.56	**6.64**	**8.93**	**9.25**	**10.02**	4.13	6.25	4.34	**9.02**	8.67	**12.47**	18.32	20.36	9.53
444M	w/o oscillation	**5.96**	**7.19**	8.37	**10.52**	10.71	**5.28**	7.29	5.13	10.36	11.00	19.29	19.58	17.19	10.61
	w/ oscillation	3.57	5.95	**8.50**	10.23	**11.38**	5.06	**8.79**	**5.95**	**19.61**	**18.33**	**21.56**	**28.58**	**24.31**	**13.22**

on general performance (Table 4). Evaluated with the lm-evaluation-harness, oscillatory models achieve lower FineWeb-Edu and LAMBADA perplexity at both scales. Average downstream accuracy is comparable at 187M parameters (43.43 vs. 43.63) and modestly higher at 444M parameters (51.08 vs. 50.18). Together, these evaluations show that oscillation benefits length generalization while maintaining general performance, with improvements on both long-context and standard downstream tasks at the larger scale.

Table 4: Standard language modeling and downstream evaluation after pretraining on 100B tokens of FineWeb-Edu. We compare Mamba-3 models with/without oscillatory dynamics at two scales.

Params	Dynamics	FW-Edu	LAMB.	LAMB.	HellaS.	PIQA	Arc-E	Arc-C	WinoGr.	OBQA	Average
		ppl↓	ppl↓	acc↑	acc_n↑	acc↑	acc↑	acc_n↑	acc↑	acc↑	acc↑
187M	w/o oscillation	16.32	38.04	31.73	40.86	**67.41**	60.73	**28.84**	**53.83**	22.00	**43.63**
	w/ oscillation	**16.18**	**32.97**	**33.59**	**41.24**	67.03	**61.28**	25.85	51.62	**23.40**	43.43
444M	w/o oscillation	12.68	18.38	41.01	52.40	70.62	69.02	32.94	**57.06**	**28.20**	50.18
	w/ oscillation	**12.60**	**17.19**	**42.83**	**52.65**	**71.65**	**70.45**	**35.15**	56.83	28.00	**51.08**

Main Findings: Oscillation improves length generalization, with greater downstream benefits at the larger model scale. Models with oscillation achieve lower long-context perplexity at both scales. Downstream scores improve at 444M and remain comparable at 187M.

5.3 The Role of Phase in Long-context Language Modeling

To understand the long-context benefits, we investigate how learned phase shapes recurrent computation in pretrained 1.5B Mamba-3. We examine phase trajectories and use interventions to study their contribution to prediction, state accumulation, and write–read alignment.

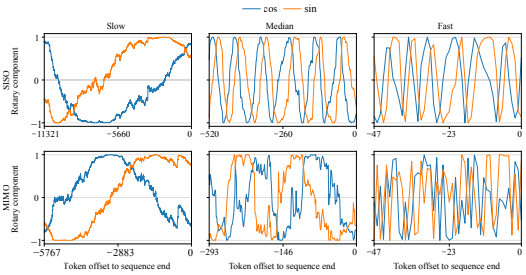

Figure 4: Representative phase-channel trajectories in pretrained Mamba-3, showing oscillations at various timescales during language processing.

Learned Phase Dynamics. Fig. 4 shows representative phase channels during natural language processing. Their trajectories exhibit oscillations at different timescales, reflecting the temporal structure of the learned rotations.

Interventions on Phase Structure. We perturb the rotation angles at each recurrent step through three interventions: *phase shuffle* permutes angles across rotary state pairs, *phase removal* sets them to zero, and *phase reversal* flips their signs. These interventions leave decay factors unchanged. For comparison, we apply *decay permutation*, which permutes decay factors across hidden states while preserving phases. We evaluate all interventions in SISO and MIMO models and give implementation details in Appendix D.4.

As shown in Table 5, all three phase interventions substantially increase PG-19 perplexity. In SISO, their effects grow strongly with context length: phase removal and reversal raise perplexity to $41.33\times$ and $74.10\times$ the native value at 64K. Phase shuffle is less damaging than decay permutation at 2K but more damaging at 64K. In MIMO, phase interventions have weaker length dependence but consistently cause greater degradation than decay permutation. These results show that long-context prediction heavily relies on the learned phase structure and the assignment to state pairs.

Phase and Recurrent State Accumulation. We examine how phase affects the combination of retained history and incoming updates using two statistics: their cosine similarity and the RMS norm of the recurrent state. Cosine similarity measures directional alignment, with larger values favoring constructive accumulation, while the RMS norm measures the resulting state magnitude.

Table 6: Hidden-state statistics under native phase and phase removal on PG-19.

Model	Length	Cosine Similarity		RMS Norm	
		Native	Remove	Native	Remove
SISO	16K	0.293	0.429	7.635	8.899
	64K	0.294	0.474	7.621	9.437
MIMO	16K	0.241	0.488	10.073	12.022
	64K	0.242	0.499	10.078	12.047

As shown in Table 6, removing phase increases both history–update alignment and state magnitude at 16K and 64K in SISO and MIMO models. These observations suggest that learned rotations moderate constructive accumulation by reorienting retained information relative to new updates. Previous work links excessive state-norm growth to poor length extrapolation in Mamba-based models (Lu et al., 2025; Shen et al., 2026), making this regulation a possible contributor to the

Table 5: PG-19 perplexity under phase interventions. Higher indicates more damaging.

Model	Intervention	2K	16K	64K
SISO	Native	21.09	21.41	20.60
	Decay permutation	114.52 (5.43×)	126.51 (5.91×)	125.27 (6.08×)
	Phase shuffle	50.24 (2.38×)	96.05 (4.49×)	149.37 (7.25×)
	Phase removal	67.99 (3.22×)	285.66 (13.34×)	851.49 (41.33×)
	Phase reversal	**152.27 (7.22×)**	**597.91 (27.93×)**	**1526.52 (74.10×)**
MIMO	Native	21.12	21.47	20.52
	Decay permutation	25.87 (1.22×)	25.94 (1.21×)	24.89 (1.21×)
	Phase shuffle	64.43 (3.05×)	70.44 (3.28×)	65.72 (3.20×)
	Phase removal	**197.17 (9.34×)**	**320.40 (14.92×)**	**309.36 (15.08×)**
	Phase reversal	74.34 (3.52×)	78.76 (3.67×)	73.89 (3.60×)

long-context benefits of oscillation. Following Shen et al. (2026), we further analyze how rotation affects state accumulation in Appendix C.3.

Notice that the effect of removing phase from pretrained models differs from training without oscillation. Our non-oscillatory models trained from scratch also maintain stable perplexity beyond the training context, without the severe degradation (Sec. 5.2). Oscillation therefore improves long-context performance in our experiments but is not required for stable length extrapolation. Understanding how models trained without oscillation achieve this stability remains an open question.

Phase Consistency Between Writing and Readout. We test whether writing and readout depend on compatible phase coordinates by removing phase rotations from B (writing), C (readout), or both. Interventions apply across all layers during the final 100 tokens of 64K windows, while the preceding context is processed with native phase.

Table 7: PG-19 perplexity under phase removal during the final 100 tokens.

Intervention	SISO	MIMO
Native	20.60	20.52
Remove B	186.28 (9.04×)	100.41 (4.89×)
Remove C	250.90 (12.18×)	117.68 (5.74×)
Remove both	42.79 (2.08×)	59.46 (2.90×)

Table 7 shows that removing rotation from either side alone substantially increases perplexity, whereas removing it from both sides partially restores performance. This recovery supports the importance of phase consistency between writing and readout. The remaining gap between joint removal and the native model may also reflect a mismatch with the state accumulated under native rotations during the preceding context.

Main Findings: Learned phase regulates how information accumulates and is accessed over long contexts. Disrupting phase degrades prediction, while removing rotation increases alignment and accumulation of old and new information. Consistent phase between writing and readout also matters for accurate prediction.

6 CONCLUSION

We interpret Mamba-3's data-dependent rotations as underdamped second-order dynamics, connecting its recurrent structure to oscillatory computation. Controlled experiments show that underdamping benefits state tracking and that oscillation improves long-context prediction and length generalization. Phase interventions reveal the importance of learned rotations for long-context prediction. Hidden-state measurements suggest that rotation moderates information accumulation, while write–read interventions highlight the importance of phase consistency. These findings highlight complementary roles for decay and phase: decay controls how much information persists, while phase shapes how it combines with new inputs and is accessed over time.

Connections to Linear Attention and Recurrent Architectures. This dynamical perspective links SSMs, linear-attention models, and other recurrent architectures through their state evolution, motivating input-dependent oscillation as a design dimension for future sequence models (Appendix A).

AI USE STATEMENT

In this work, we used generative AI tools for language polishing and code generation. All AI-assisted text was reviewed and revised by the authors. AI-generated code was inspected, tested, and verified by the authors for correctness before being used in the experiments. We take full responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with the aid of generative AI.

REFERENCES

Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. In *International conference on machine learning*, pp. 1120–1128. PMLR, 2016.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 3119–3137, 2024.

Assaf Ben-Kish, Itamar Zimerman, Shady Abu-Hussein, Nadav Cohen, Amir Globerson, Lior Wolf, and Raja Giryes. Decimamba: Exploring the length extrapolation potential of mamba. In *International Conference on Learning Representations*, volume 2025, pp. 101148–101170, 2025.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.

Jared Boyer, T Konstantin Rusch, and Daniela Rus. Learning to dissipate energy in oscillatory state-space models. *arXiv preprint arXiv:2505.12171*, 2025.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL `https://zenodo.org/records/12608602`.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Riccardo Grazzi, Julien Siems, Arber Zela, Jörg Franke, Frank Hutter, et al. Unlocking state-tracking in linear rnns through negative eigenvalues. In *International Conference on Learning Representations*, volume 2025, pp. 36565–36597, 2025.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. On the parameterization and initialization of diagonal state space models. *Advances in neural information processing systems*, 35:35971–35983, 2022.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. Linearity of relation decoding in transformer language models. In *International Conference on Learning Representations*, volume 2024, pp. 10504–10526, 2024.

Yulong Huang, Xiang Liu, Hongxiang Huang, Xiaopeng Lin, Zunchang Liu, Xiaowen Chu, Zeke Xie, and Bojun Cheng. Mdn: Parallelizing stepwise momentum for delta linear attention. *arXiv preprint arXiv:2605.05838*, 2026.

Li Jing, Yichen Shen, Tena Dubcek, John Peurifoy, Scott Skirlo, Yann LeCun, Max Tegmark, and Marin Soljačić. Tunable efficient unitary neural networks (eunn) and their application to rnns. In *International Conference on Machine Learning*, pp. 1733–1741. PMLR, 2017.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5156–5165. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/katharopoulos20a.html.

T Anderson Keller and Max Welling. Neural wave machines: learning spatiotemporally structured representations with locally coupled oscillatory recurrent neural networks. In *International Conference on Machine Learning*, pp. 16168–16189. PMLR, 2023.

Aakash Sunil Lahoti, Kevin Li, Berlin Chen, Caitlin Wang, Aviv Bick, Zico Kolter, Tri Dao, and Albert Gu. Mamba-3: Improved sequence modeling using state space principles. In *International Conference on Learning Representations*, volume 2026, pp. 86173–86202, 2026.

Peng Lu, Jerry Huang, Qiuhao Zeng, Xinyu Wang, Boxing Chen, Philippe Langlais, and Yufei Cui. Mamba modulation: On the length generalization of mamba models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=QEU047bE8p.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2381–2391, 2018.

Sajad Movahedi, Timur Carstensen, Arshia Afzal, Frank Hutter, Antonio Orvieto, and Volkan Cevher. Selective rotary position embedding. In *International Conference on Learning Representations*, volume 2026, pp. 10219–10247, 2026.

Antonio Orvieto, Samuel L Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. In *International conference on machine learning*, pp. 26670–26698. PMLR, 2023.

Antonio Orvieto, Soham De, Caglar Gulcehre, Razvan Pascanu, and Samuel L Smith. Universality of linear recurrences followed by non-linear projections: Finite-width guarantees and benefits of complex eigenvalues. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 38837–38863. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/orvieto24a.html.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc-Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The lambada dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1525–1534, 2016.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, and Timothy P Lillicrap. Compressive transformers for long-range sequence modelling. *arXiv preprint arXiv:1911.05507*, 2019.

T. Konstantin Rusch and Siddhartha Mishra. Coupled oscillatory recurrent neural network (cornn): An accurate and (gradient) stable architecture for learning long time dependencies. In *International Conference on Learning Representations*, 2021.

T Konstantin Rusch and Daniela Rus. Oscillatory state-space models. In *International Conference on Learning Representations*, volume 2025, pp. 22303–22322, 2025.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *Proc. Int. Conf. on Machine Learning (ICML)*, Virtual only, July 2021.

Arnab Sen Sharma, David Atkinson, and David Bau. Locating and editing factual associations in mamba. *arXiv preprint arXiv:2404.03646*, 2024.

Haochen Shen, Davis Wertheimer, Zheng Wang, Garrett Goon, Derrick Liu, Naigang Wang, Mudhakar Srivatsa, Raghu Ganti, and Minjia Zhang. From collapse to control: Understanding and extending context length in emerging hybrid models via universal position interpolation. In *International Conference on Learning Representations*, volume 2026, pp. 122566–122589, 2026.

Julien Siems, Riccardo Grazzi, Korbinian Pöppel, Jaisidh Singh, Arber Zela, Timur Carstensen, Jenia Jitsev, Frank Hutter, Volkan Cevher, Antonio Orvieto, and Aaron Klein. Complex kda: Understanding and enhancing the expressivity of kimi delta attention, 2026. URL `https://arxiv.org/abs/2609.24797`.

Jimmy T.H. Smith, Andrew Warrington, and Scott Linderman. Simplified state space layers for sequence modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=Ai8Hw3AXqks`.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama, June 2023. URL `https://huggingface.co/datasets/cerebras/SlimPajama-627B`.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Kimi Team, Yu Zhang, Zongyu Lin, Xingcheng Yao, Jiaxi Hu, Fanqing Meng, Chengyin Liu, Xin Men, Songlin Yang, Zhiyuan Li, Wentao Li, Enzhe Lu, Weizhou Liu, Yanru Chen, Weixin Xu, Longhui Yu, Yejie Wang, Yu Fan, Longguang Zhong, Enming Yuan, Dehao Zhang, Yizhi Zhang, T. Y. Liu, Haiming Wang, Shengjun Fang, Weiran He, Shaowei Liu, Yiwei Li, Jianlin Su, Jiezhong Qiu, Bo Pang, Junjie Yan, Zhejun Jiang, Weixiao Huang, Bohong Yin, Jiacheng You, Chu Wei, Zhengtao Wang, Chao Hong, Yutian Chen, Guanduo Chen, Yucheng Wang, Huabin Zheng, Feng Wang, Yibo Liu, Mengnan Dong, Zheng Zhang, Siyuan Pan, Wenhao Wu, Yuhao Wu, Longyu Guan, Jiawen Tao, Guohong Fu, Xinran Xu, Yuzhi Wang, Guokun Lai, Yuxin Wu, Xinyu Zhou, Zhilin Yang, and Yulun Du. Kimi linear: An expressive, efficient attention architecture, 2025. URL `https://arxiv.org/abs/2510.26692`.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. *Natural language processing with transformers*. O'Reilly Media, Inc., 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Scott Wisdom, Thomas Powers, John Hershey, Jonathan Le Roux, and Les Atlas. Full-capacity unitary recurrent neural networks. *Advances in neural information processing systems*, 29, 2016.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. *Advances in neural information processing systems*, 37:115491–115522, 2024.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=r8H7xhYPwz.

Zhifan Ye, Kejing Xia, Yonggan Fu, Xin Dong, Jihoon Hong, Xiangchi Yuan, Shizhe Diao, Jan Kautz, Pavlo Molchanov, and Yingyan Celine Lin. Longmamba: Enhancing mamba's long-context capabilities via training-free receptive field enlargement. In *International Conference on Learning Representations*, volume 2025, pp. 68486–68504, 2025.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 2019.

APPENDIX CONTENTS

A Extension to Linear Attention and Recurrent Models **15**

A.1 Homogeneous state propagation 15

A.2 Linear Attention 16

A.3 DeltaNet 16

A.4 Gated DeltaNet 17

A.5 Kimi Delta Attention 18

A.6 Complex KDA 19

A.7 Momentum DeltaNet 20

B Related Work **21**

C Mathematical Derivations **21**

C.1 Matrix Exponential of the Rotation Generator 21

C.2 Mode Persistence Across Damping Regimes 22

C.3 State Accumulation Under Persistent Forcing 23

C.4 Householder Reflections for Rotation 24

D Implementation Details **24**

D.1 State Tracking Tasks 24

D.2 Oscillation across Language Tasks 25

D.3 From-Scratch Pretraining 25

D.4 Pretrained Model Phase Intervention 25

D.5 LongBench-E Task Abbreviations 27

E Additional Results **27**

E.1 Phase Structure Induced by Oscillation 27

E.2 Length Generalization on Other Datasets 28

E.3 Training Dynamics 28

A Extension to Linear Attention and Recurrent Models

Although this work primarily focuses on Mamba-3, the dynamical-systems perspective also extends to the homogeneous transitions of other linear-attention models. Table 8 summarizes representative update rules and their homogeneous dynamics in an interval. Complex KDA can admit oscillatory modes, while Momentum DeltaNet (MDN) explicitly introduces a momentum state that permits damped oscillations (Siems et al., 2026; Huang et al., 2026). **How to guarantee and control underdamped modes within linear-attention updates remains an interesting direction for future work.** Below, we characterize their spectral regimes and derive continuous-time generators where they exist.

Table 8: Linear attention models and their homogeneous dynamics within an interval. Delta-rule models use unit-norm keys. For DeltaNet, Gated DeltaNet, and KDA, we assume $\beta_t \in (0, 1)$; the decay gates in Gated DeltaNet and KDA lie in $(0, 1)$. Complex KDA allows $\beta_t \in [0, 2]$ and $\alpha_{t,i} \in [-1, 1]$. MDN uses its constrained gate parameterization.

Model	Update rule	Dynamics
Linear Attention (Katharopoulos et al., 2020)	$S_t = S_{t-1} + k_t v_t^\top$	No Decay
DeltaNet (Schlag et al., 2021; Yang et al., 2024)	$S_t = H_t S_{t-1} + \beta_t k_t v_t^\top$	Decay with a preserved mode
Gated DeltaNet (Yang et al., 2025)	$S_t = \alpha_t H_t S_{t-1} + \beta_t k_t v_t^\top$	Non-oscillatory two-rate decay
KDA (Team et al., 2025; 2026)	$S_t = H_t D_t S_{t-1} + \beta_t k_t v_t^\top$	Non-oscillatory modal decay
Complex KDA (Siems et al., 2026)	$S_t = H_t D_t S_{t-1} + \beta_t k_t v_t^\top$	Can admit oscillatory modes
MDN (Huang et al., 2026)	$M_t = \mu_t M_{t-1} - \eta_t k_t e_t^\top$, $S_t = \alpha_t S_{t-1} - \beta_t M_t$	Can admit oscillatory modes

Table 9: Notation and definitions.

Notation	Meaning
I	Identity matrix
S_t	Recurrent state matrix
q_t	Query vector
k_t	Key vector
v_t	Value vector
o_t	Readout vector, $S_t^\top q_t$
T_t	State transition matrix
U_t	Additive input term
β_t	Input-dependent delta-rule update gate
H_t	Delta-rule transition matrix, $I - \beta_t k_t k_t^\top$
α_t	Scalar decay gate
$\boldsymbol{\alpha}_t$	Vector of per-channel gates
D_t	Diagonal gate matrix, $\mathrm{Diag}(\boldsymbol{\alpha}_t)$
M_t	Momentum state matrix in MDN
μ_t	Momentum retention gate in MDN
η_t	Correction scale in MDN
e_t	Prediction error, $v_t - \alpha_t S_{t-1}^\top k_t$

A.1 Homogeneous state propagation

Following Sec. 3, we analyze homogeneous state propagation with coefficients held fixed within each interval. For the matrix-valued recurrence

$$S_t = T_t S_{t-1} + U_t, \qquad o_t = S_t^\top q_t, \tag{28}$$

we omit the additive term U_t and examine the transition T_t, which acts on each column $s_{t-1} \in \mathbb{R}^{d_k}$ of S_{t-1}. For MDN, we instead analyze the block transition of the augmented state (S_t, M_t). Within an interval of length $\Delta > 0$, we suppress the index t and construct continuous-time realizations of the frozen transitions.

For each selected invariant two-dimensional subsystem of the single-state models, we use $x = (u, v)^\top$ to denote its two scalar state coordinates. We specify these coordinates for each transition below and, when they are coupled, eliminate v from the corresponding first-order dynamics to obtain a second-order equation for u. Where the resulting equation has $\omega_0 > 0$, we compare it with the canonical form

$$\ddot{u} + 2\zeta\omega_0\dot{u} + \omega_0^2 u = 0. \tag{29}$$

Modes with a zero continuous-time eigenvalue, such as DeltaNet's preserved mode, are described directly rather than assigned a finite damping ratio. MDN has an additional momentum state, so we analyze its augmented state (S_t, M_t) separately below. Here u and v are scalar coordinates of the selected subsystem; they need not be entries of the original state column s. All derivatives are taken within the fixed interval.

A.2 Linear Attention

Linear Attention has $T = I$. Its homogeneous update is reproduced by $\dot{S} = 0$, so the state is preserved within the interval. The full recurrence accumulates input contributions,

$$S_t = S_0 + \sum_{j=1}^{t} U_j. \tag{30}$$

A.3 DeltaNet

Let $k \in \mathbb{R}^{d_k}$ be a unit-norm key, with $d_k \geq 2$, and define

$$\Pi = kk^\top, \qquad \Pi^2 = \Pi. \tag{31}$$

For $0 < \beta < 1$, the homogeneous transition is

$$T = I - \beta\Pi = (I - \Pi) + (1 - \beta)\Pi. \tag{32}$$

Thus, the component along k is multiplied by $1 - \beta$, while every component orthogonal to k is preserved. Define

$$b = -\frac{\log(1 - \beta)}{\Delta} > 0, \qquad A = -b\Pi. \tag{33}$$

Since Π is a projector, $\Pi^n = \Pi$ for every integer $n \geq 1$. Expanding the matrix exponential gives

$$\begin{aligned} e^{A\Delta} &= \sum_{n=0}^{\infty} \frac{(A\Delta)^n}{n!} \\ &= I + \left(\sum_{n=1}^{\infty} \frac{(-b\Delta)^n}{n!} \right) \Pi \\ &= I + \left(e^{-b\Delta} - 1\right) \Pi \\ &= I - \beta\Pi = T. \end{aligned} \tag{34}$$

Hence $\dot{s} = As$ reproduces the discrete homogeneous transition over the interval. Choose an orthonormal basis $\{k, w_1, \ldots, w_{d_k - 1}\}$ satisfying

$$k^\top w_i = 0, \qquad w_i^\top w_j = \delta_{ij}. \tag{35}$$

The state admits the decomposition

$$s = \mu k + \sum_{i=1}^{d_k - 1} \nu_i w_i, \qquad \mu = k^\top s, \qquad \nu_i = w_i^\top s. \tag{36}$$

These scalar coordinates are projections onto the chosen basis directions rather than necessarily entries of s in its original coordinate system. Using $A = -bkk^\top$, together with $k^\top k = 1$ and $k^\top w_i = 0$, we have

$$\begin{aligned} Ak &= -bkk^\top k = -bk(k^\top k) = -bk, \\ Aw_i &= -bkk^\top w_i = -bk(k^\top w_i) = 0. \end{aligned} \tag{37}$$

Thus, each plane $\mathcal{E}_i = \mathrm{span}\{k, w_i\}$ is invariant. On this plane, the modal coordinates satisfy

$$\dot{\mu} = k^\top \dot{s} = k^\top As = -bk^\top kk^\top s = -b(k^\top k)(k^\top s) = -b\mu, \tag{38}$$

Similarly,

$$\dot{\nu}_i = w_i^\top \dot{s} = w_i^\top As = -bw_i^\top kk^\top s = -b(w_i^\top k)(k^\top s) = 0, \tag{39}$$

To obtain an equivalent second-order representation on $\mathcal{E}_i$, select one plane $\mathcal{E}_i$ and introduce the invertible coordinates

$$u = \mu + \nu_i, \quad v = \mu, \quad x = (u, v)^\top, \tag{40}$$

so (u, v) retains the full two-dimensional state on $\mathcal{E}_i$. Their dynamics are

$$\dot{u} = \dot{\mu} + \dot{\nu}_i = -bv, \quad \dot{v} = -bv, \tag{41}$$

Hence

$$\ddot{u} = -b\dot{v} = b^2 v = -b\dot{u}. \tag{42}$$

Therefore,

$$\ddot{u} + b\dot{u} = 0. \tag{43}$$

On each plane $\mathcal{E}_i$, the eigenvalues are $-b$ and 0, corresponding to a decaying and a preserved mode, respectively. In the full state space, however, $-b$ has multiplicity one, while 0 has multiplicity $d_k - 1$. Thus, although every orthogonal direction w_i can be represented by a second-order representation, all such planes share the same decaying mode along k (the coordinate μ, or v after the transformation). **They are overlapping representations of the full system, rather than independent two-dimensional subsystems. The same shared-direction structure persists in Gated DeltaNet.**

A.4 Gated DeltaNet

Gated DeltaNet uses $T = \alpha H$ with $0 < \alpha < 1$ (Yang et al., 2025). With b and Π defined above, let

$$\gamma = -\frac{\log \alpha}{\Delta} > 0, \quad A = -\gamma I - b\Pi. \tag{44}$$

Since I and Π commute,

$$\begin{aligned} e^{A\Delta} &= e^{-\gamma\Delta}\left[I + (e^{-b\Delta} - 1)\Pi\right] \\ &= \alpha(I - \beta\Pi) = \alpha H = T. \end{aligned} \tag{45}$$

Hence $\dot{s} = As$ reproduces the discrete homogeneous transition over the interval.

Use the orthonormal basis $\{k, w_1, \ldots, w_{d_k-1}\}$ defined in the DeltaNet subsection. The action of A on its basis vectors is

$$\begin{aligned} Ak &= -\gamma k - bk(k^\top k) = -(\gamma + b)k, \\ Aw_i &= -\gamma w_i - bk(k^\top w_i) = -\gamma w_i. \end{aligned} \tag{46}$$

Thus each plane $\mathcal{E}_i = \mathrm{span}\{k, w_i\}$ is invariant. Its modal coordinates $\mu = k^\top s$ and $\nu_i = w_i^\top s$ satisfy

$$\begin{aligned} \dot{\mu} &= k^\top As = -(\gamma + b)\mu, \\ \dot{\nu}_i &= w_i^\top As = -\gamma\nu_i. \end{aligned} \tag{47}$$

The key-direction coordinate μ decays at rate $\gamma + b$, while every orthogonal coordinate ν_i decays at rate γ.

As in DeltaNet, select one plane $\mathcal{E}_i$ and define

$$u = \mu + \nu_i, \quad v = \mu, \quad x = (u, v)^\top. \tag{48}$$

This coordinate change is invertible on $\mathcal{E}_i$, with $\mu = v$ and $\nu_i = u - v$. The dynamics become

$$\begin{aligned} \dot{u} &= \dot{\mu} + \dot{\nu}_i = -(\gamma + b)v - \gamma(u - v) = -\gamma u - bv, \\ \dot{v} &= -(\gamma + b)v. \end{aligned} \tag{49}$$

Differentiating $\dot{u}$ and using $bv = -(\dot{u} + \gamma u)$ gives

$$\begin{aligned} \ddot{u} &= -\gamma\dot{u} - b\dot{v} \\ &= -\gamma\dot{u} + b(\gamma + b)v \\ &= -\gamma\dot{u} - (\gamma + b)(\dot{u} + \gamma u) \\ &= -(2\gamma + b)\dot{u} - \gamma(\gamma + b)u. \end{aligned} \tag{50}$$

Therefore,

$$\ddot{u} + (2\gamma + b)\dot{u} + \gamma(\gamma + b)u = 0. \tag{51}$$

The canonical parameters are

$$\omega_0 = \sqrt{\gamma(\gamma + b)}, \quad \zeta = \frac{2\gamma + b}{2\sqrt{\gamma(\gamma + b)}} > 1. \tag{52}$$

On each plane $\mathcal{E}_i$, the characteristic roots are $-(\gamma + b)$ and $-\gamma$, so the subsystem is overdamped. In the full state space, $-(\gamma + b)$ has multiplicity one, while $-\gamma$ has multiplicity $d_k - 1$. As in DeltaNet, each ν_i can be paired with μ, **but all such planes share the single key-direction mode**. Their second-order representations therefore overlap.

A.5 Kimi Delta Attention

KDA replaces the scalar gate with a diagonal gate, $T = HD$, where $D = \mathrm{Diag}(\boldsymbol{\alpha})$ (Team et al., 2025; 2026). Assume $0 < \beta < 1$ and $0 < \alpha_i < 1$. Although $T = HD$ need not be symmetric, the similarity transform

$$\widehat{T} = D^{1/2} T D^{-1/2} = D^{1/2} H D^{1/2} \tag{53}$$

is symmetric. Since $(1 - \beta)I \preceq H \preceq I$, it satisfies

$$0 \prec (1 - \beta)D \preceq \widehat{T} \preceq D \prec I. \tag{54}$$

Therefore T has a real eigenbasis and eigenvalues $\lambda_\ell \in (0, 1)$. Write

$$T = Q\,\mathrm{Diag}(\lambda_1, \ldots, \lambda_{d_k}) Q^{-1}, \tag{55}$$

where the columns q_ℓ of Q are eigenvectors of T. For $\Delta > 0$, define

$$b_\ell = -\frac{\log \lambda_\ell}{\Delta} > 0, \quad A = Q\,\mathrm{Diag}(-b_1, \ldots, -b_{d_k}) Q^{-1}. \tag{56}$$

Then

$$e^{A\Delta} = Q\,\mathrm{Diag}(e^{-b_1\Delta}, \ldots, e^{-b_{d_k}\Delta}) Q^{-1} = T. \tag{57}$$

For a state column s, define its modal coordinates by

$$c = Q^{-1} s, \quad s = \sum_{\ell=1}^{d_k} c_\ell q_\ell. \tag{58}$$

These are expansion coefficients in the eigenvector basis, which need not be orthogonal. Since $s = Qc$ and Q is fixed within the frozen interval, $\dot{s} = Q\dot{c}$. Using $\dot{s} = As$, we obtain

$$\dot{c} = Q^{-1} A Q c = -\,\mathrm{Diag}(b_1, \ldots, b_{d_k}) c. \tag{59}$$

Thus $\dot{c}_\ell = -b_\ell c_\ell$. The rates b_ℓ describe eigenmodes of $T = HD$.

Select two modes with distinct rates $b_i \neq b_j$. Their eigenvectors span an invariant plane $\mathcal{E}_{ij} = \mathrm{span}\{q_i, q_j\}$. Let

$$\mu = c_i = (Q^{-1} s)_i, \quad \nu = c_j = (Q^{-1} s)_j. \tag{60}$$

On this plane,

$$\dot{\mu} = -b_i \mu, \quad \dot{\nu} = -b_j \nu. \tag{61}$$

As in the preceding subsections, define

$$u = \mu + \nu, \quad v = \mu, \quad x = (u, v)^\top. \tag{62}$$

This coordinate change is invertible on $\mathcal{E}_{ij}$, with $\mu = v$ and $\nu = u - v$. The dynamics become

$$\begin{aligned} \dot{u} = \dot{\mu} + \dot{\nu} = -b_i v - b_j(u - v) = -b_j u - (b_i - b_j)v, \\ \dot{v} = -b_i v. \end{aligned} \tag{63}$$

Differentiating $\dot{u}$ and using $(b_i - b_j)v = -(\dot{u} + b_j u)$ gives

$$\begin{aligned} \ddot{u} &= -b_j \dot{u} - (b_i - b_j)\dot{v} \\ &= -b_j \dot{u} + b_i(b_i - b_j)v \\ &= -b_j \dot{u} - b_i(\dot{u} + b_j u) \\ &= -(b_i + b_j)\dot{u} - b_i b_j u. \end{aligned} \tag{64}$$

Therefore,

$$\ddot{u} + (b_i + b_j)\dot{u} + b_i b_j u = 0. \tag{65}$$

Its canonical parameters are

$$\omega_0 = \sqrt{b_i b_j}, \quad \zeta = \frac{b_i + b_j}{2\sqrt{b_i b_j}} > 1. \tag{66}$$

The characteristic roots are $-b_i$ and $-b_j$; hence, for $b_i \neq b_j$, the subsystem is overdamped. Since KDA can exhibit multiple distinct modal decay rates, disjoint eigenmode pairs can **form non-overlapping invariant subspaces that evolve independently in modal coordinates.** This differs from DeltaNet and Gated DeltaNet, where all such two-dimensional pairings share the same key-direction mode.

A.6 Complex KDA

Complex KDA (Siems et al., 2026) allows $\beta \in [0, 2]$ and $\alpha_i \in [-1, 1]$. Its frozen homogeneous transition is

$$T = (I - \beta k k^\top)D, \quad D = \mathrm{Diag}(\boldsymbol{\alpha}), \tag{67}$$

To exhibit an oscillatory mode arising from signed channel-wise gating, consider the two-dimensional construction in (Siems et al., 2026). Let

$$H_k = I - 2kk^\top, \quad T_{\alpha,\theta} = H_k D_\alpha, \quad D_\alpha = \mathrm{Diag}(\alpha, 1), \quad k = (\cos(\theta/2), \sin(\theta/2))^\top. \tag{68}$$

As shown in Appendix C.4, expanding the product gives

$$T_{\alpha,\theta} = \begin{bmatrix} -\alpha\cos\theta & -\sin\theta \\ -\alpha\sin\theta & \cos\theta \end{bmatrix}. \tag{69}$$

Its discriminant is

$$\mathcal{D} = [\mathrm{tr}(T_{\alpha,\theta})]^2 - 4\det(T_{\alpha,\theta}) = (1 - \alpha)^2\cos^2\theta + 4\alpha. \tag{70}$$

A non-real conjugate pair occurs precisely when

$$\alpha < 0 \quad \text{and} \quad |\cos\theta| < \frac{2\sqrt{-\alpha}}{1 - \alpha}. \tag{71}$$

Under this condition, the eigenvalues are

$$\lambda_\pm = \frac{(1 - \alpha)\cos\theta}{2} \pm \frac{i}{2}\sqrt{-\mathcal{D}}. \tag{72}$$

Since $\lambda_+\lambda_- = \det(T_{\alpha,\theta}) = -\alpha$, their common modulus is $\rho = \sqrt{-\alpha}$. For $-1 < \alpha < 0$, we have $0 < \rho < 1$ and may write

$$\lambda_\pm = \rho e^{\pm i\phi}, \qquad \phi = \arccos\left(\frac{(1 - \alpha)\cos\theta}{2\rho}\right) \in (0, \pi). \tag{73}$$

For an interval of length $\Delta > 0$, define

$$a = -\frac{\log\rho}{\Delta} > 0, \qquad \Omega = \frac{\phi}{\Delta} > 0. \tag{74}$$

The real matrix $T_{\alpha,\theta}$ is similar to $\rho R(\phi)$. In the corresponding real modal coordinates $x = (u, v)^\top$, a corresponding continuous-time realization is

$$\dot{x} = \begin{bmatrix} -a & -\Omega \\ \Omega & -a \end{bmatrix} x, \qquad e^{(-aI+\Omega J)\Delta} = \rho R(\phi). \tag{75}$$

As in the decay–rotation analysis of Mamba-3 in Sec. 3, eliminating v yields

$$\ddot{u} + 2a\dot{u} + (a^2 + \Omega^2)u = 0, \tag{76}$$

with $\omega_0 = \sqrt{a^2 + \Omega^2}$ and $\zeta = a/\omega_0 < 1$. Thus, under these conditions, **this CKDA mode admits an underdamped second-order realization.**

A.7 Momentum DeltaNet

Momentum DeltaNet (MDN) augments the recurrent state S_t with a momentum state M_t of the same shape (Huang et al., 2026). Its update is

$$\begin{aligned} M_t &= \mu_t M_{t-1} - \eta_t k_t \big(v_t - \alpha_t S_{t-1}^\top k_t\big)^\top, \\ S_t &= \alpha_t S_{t-1} - \beta_t M_t. \end{aligned} \tag{77}$$

Freeze the coefficients and a unit-norm key k within an interval of length $\Delta > 0$, and suppress the index t. To study homogeneous propagation, omit the external value input v_t. With $\Pi = kk^\top$, the coupled update becomes

$$\begin{bmatrix} S_t \\ M_t \end{bmatrix} = \underbrace{\begin{bmatrix} \alpha(I - \beta\eta\Pi) & -\beta\mu I \\ \alpha\eta\Pi & \mu I \end{bmatrix}}_{\mathcal{T}} \begin{bmatrix} S_{t-1} \\ M_{t-1} \end{bmatrix}. \tag{78}$$

Consider corresponding columns s and m of S and M. Their key-direction coordinates

$$u = k^\top s, \qquad p = k^\top m \tag{79}$$

form an invariant two-dimensional subsystem. Since $\Pi k = k$, its transition is

$$\begin{bmatrix} u_t \\ p_t \end{bmatrix} = B \begin{bmatrix} u_{t-1} \\ p_{t-1} \end{bmatrix}, \qquad B = \begin{bmatrix} \alpha(1 - \beta\eta) & -\beta\mu \\ \alpha\eta & \mu \end{bmatrix}. \tag{80}$$

The trace and determinant of B are

$$\tau = \alpha(1 - \beta\eta) + \mu, \qquad \det(B) = \alpha\mu. \tag{81}$$

Hence its characteristic discriminant is

$$\mathcal{D} = \tau^2 - 4\alpha\mu. \tag{82}$$

When $\mathcal{D} < 0$, the eigenvalues form a non-real conjugate pair. For $0 < \alpha, \mu < 1$, define

$$\rho = \sqrt{\alpha\mu} \in (0, 1), \qquad \phi = \arccos\left(\frac{\tau}{2\rho}\right) \in (0, \pi). \tag{83}$$

When $\mathcal{D} < 0$, the eigenvalues are $\lambda_\pm = \rho e^{\pm i\phi}$. Set

$$K = \frac{B - \rho\cos\phi\, I}{\rho\sin\phi}. \tag{84}$$

The characteristic equation of B gives $K^2 = -I$, and thus

$$B = \rho(\cos\phi\, I + \sin\phi\, K). \tag{85}$$

Define

$$a = -\frac{\log\rho}{\Delta} > 0, \qquad \Omega = \frac{\phi}{\Delta} > 0, \qquad A = -aI + \Omega K. \tag{86}$$

Since $K^2 = -I$, we have $e^{A\Delta} = B$. Moreover, K is similar over the reals to the rotation generator J in Sec. 3. Thus, when $\mathcal{D} < 0$ and $0 < \alpha, \mu < 1$, this frozen key-direction subsystem admits an underdamped second-order realization.

B Related Work

Oscillatory Dynamics and Rotations in Recurrent Models. Oscillatory systems provide a dynamical foundation for recurrent architectures. CoRNN (Rusch & Mishra, 2021) discretizes coupled nonlinear oscillators and analyzes the stability of states and gradients, while Neural Wave Machines (Keller & Welling, 2023) use locally coupled oscillators to learn spatiotemporally structured representations. LinOSS (Rusch & Rus, 2025) builds state-space models from forced harmonic oscillators, and D-LinOSS (Boyer et al., 2025) introduces learnable dissipation to control their damping. Selective RoPE (Movahedi et al., 2026) incorporates input-dependent rotations into selectively decaying linear recurrences, improving state tracking and language modeling. Our work connects Mamba-3 to this family of dynamical models through an underdamped second-order interpretation of its decay–rotation transition. This connection enables controlled comparisons across damping regimes and an analysis of how learned phase shapes recurrent state accumulation.

Complex Parameterizations and Spectral Structure in Recurrent Models. The spectrum of a recurrent transition governs how states persist and evolve. Early unitary recurrent networks use complex-valued states and norm-preserving transitions to facilitate learning long-term dependencies (Arjovsky et al., 2016; Wisdom et al., 2016; Jing et al., 2017). Complex parameterizations also underpin state-space and linear recurrent architectures such as S4 (Gu et al., 2021), S4D (Gu et al., 2022), S5 (Smith et al., 2023), and LRU (Orvieto et al., 2023). Complementing these architectural developments, Orvieto et al. (2024) study how complex eigenvalues support long-range information retention, and Grazzi et al. (2025) show that allowing negative eigenvalues overcomes limitations of positive-spectrum recurrences on state tracking and improves length generalization. We examine Mamba-3's complex-conjugate modes through their damping and phase dynamics, linking spectral structure to controlled training results and phase interventions in pretrained language models.

C Mathematical Derivations

C.1 Matrix Exponential of the Rotation Generator

We provide the derivation of the matrix exponential used in Section 3. Let

$$J = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \tag{87}$$

which satisfies

$$J^2 = -I. \tag{88}$$

It follows that

$$J^{2k} = (-1)^k I, \qquad J^{2k+1} = (-1)^k J. \tag{89}$$

Using the Taylor expansion of the matrix exponential,

$$e^{\phi J} = \sum_{k=0}^{\infty} \frac{(\phi J)^k}{k!}, \tag{90}$$

and separating the even and odd powers gives

$$e^{\phi J} = \sum_{k=0}^{\infty} \frac{\phi^{2k} J^{2k}}{(2k)!} + \sum_{k=0}^{\infty} \frac{\phi^{2k+1} J^{2k+1}}{(2k+1)!} \tag{91}$$

$$= \left[\sum_{k=0}^{\infty} \frac{(-1)^k \phi^{2k}}{(2k)!}\right] I + \left[\sum_{k=0}^{\infty} \frac{(-1)^k \phi^{2k+1}}{(2k+1)!}\right] J \tag{92}$$

$$= \cos(\phi) I + \sin(\phi) J. \tag{93}$$

Substituting the definition of J yields

$$e^{\phi J} = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix} = R(\phi). \tag{94}$$

For the generator

$$A^{\mathrm{rot}} = -aI + \nu J, \tag{95}$$

the matrices I and J commute. Hence,

$$\begin{aligned} e^{A^{\mathrm{rot}}\Delta} &= e^{(-aI+\nu J)\Delta} && (96)\\ &= e^{-a\Delta I} e^{\nu\Delta J} && (97)\\ &= e^{-a\Delta}\left[\cos(\nu\Delta)I + \sin(\nu\Delta)J\right] && (98)\\ &= e^{-a\Delta} R(\nu\Delta). && (99) \end{aligned}$$

With $a = -\log\rho/\Delta$ and $\nu = \theta/\Delta$, this becomes

$$e^{A^{\mathrm{rot}}\Delta} = \rho R(\theta), \tag{100}$$

which establishes the decay–rotation transition used in the main text.

C.2 Mode Persistence Across Damping Regimes

We compare homogeneous second-order dynamics at a fixed natural frequency $\omega_0 > 0$:

$$\dot{s} = A(a)s, \qquad A(a) = \begin{bmatrix} 0 & 1 \\ -\omega_0^2 & -2a \end{bmatrix}. \tag{101}$$

The eigenvalues are

$$\lambda_\pm = -a \pm \sqrt{a^2 - \omega_0^2}. \tag{102}$$

The discrete-step statements in Sec. 4.2 follow by evaluating the continuous-time flow at $t = n\Delta$.

Overdamped dynamics. For $a > \omega_0$, let

$$r = \sqrt{a^2 - \omega_0^2}. \tag{103}$$

The eigenvalues are $\lambda_\pm = -a \pm r$, where λ_+ is the slower mode. Their relative magnitude satisfies

$$\lim_{t\to\infty} \left| \frac{e^{\lambda_- t}}{e^{\lambda_+ t}} \right| = \lim_{t\to\infty} e^{-2rt} = 0. \tag{104}$$

Thus, for initial states with a nonzero slow-mode component, the fast mode becomes negligible and propagation is asymptotically dominated by the slow eigendirection. The dominant exponential decay rate is

$$a - r = a - \sqrt{a^2 - \omega_0^2}. \tag{105}$$

Critically damped dynamics. At $a = \omega_0$, write

$$A = -\omega_0 I + N, \qquad N = \begin{bmatrix} \omega_0 & 1 \\ -\omega_0^2 & -\omega_0 \end{bmatrix}. \tag{106}$$

Since $N^2 = 0$,

$$e^{At} = e^{-\omega_0 t}(I + tN). \tag{107}$$

The polynomial factor does not change the asymptotic exponential decay rate, which is therefore ω_0. Moreover,

$$\lim_{t\to\infty} \frac{e^{\omega_0 t}}{t} e^{At} = \lim_{t\to\infty} \left(\frac{I}{t} + N \right) = N. \tag{108}$$

Since

$$N = \begin{bmatrix} 1 \\ -\omega_0 \end{bmatrix} \begin{bmatrix} \omega_0 & 1 \end{bmatrix}, \tag{109}$$

we have $\mathrm{rank}(N) = 1$. Thus, after removing the common exponential decay and polynomial factor, the propagation operator becomes asymptotically rank one.

Underdamped dynamics. For $0 < a < \omega_0$, let

$$\nu = \sqrt{\omega_0^2 - a^2}. \tag{110}$$

The eigenvalues are $\lambda_\pm = -a \pm i\nu$, and hence

$$\left| \frac{e^{\lambda_- t}}{e^{\lambda_+ t}} \right| = 1 \qquad \text{for all } t \geq 0. \tag{111}$$

Thus, neither mode is preferentially attenuated. Moreover, there exists an invertible change of coordinates T such that

$$TA(a)T^{-1} = -aI + \nu J, \qquad J = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}. \tag{112}$$

Therefore,

$$Te^{A(a)t}T^{-1} = e^{-at}R(\nu t). \tag{113}$$

In these rotation coordinates, both state directions decay with the same radial factor while their relative two-dimensional geometry is preserved. The exponential decay rate is a.

Comparison. At fixed ω_0, the underdamped decay rate satisfies $a < \omega_0$. For the overdamped regime,

$$a - \sqrt{a^2 - \omega_0^2} = \frac{\omega_0^2}{a + \sqrt{a^2 - \omega_0^2}} < \omega_0, \tag{114}$$

whereas the critically damped rate is exactly ω_0. Critical damping therefore has the fastest asymptotic exponential attenuation at fixed natural frequency.

More importantly, the regimes differ in their modal geometry: underdamped propagation preserves two equally decaying modes, whereas overdamped propagation suppresses the fast mode and critical propagation becomes asymptotically rank one.

C.3 State Accumulation Under Persistent Forcing

Prior work has identified recurrent-state growth as a failure mode of Mamba under context-length extrapolation (Shen et al., 2026). Consider the scalar recurrence

$$h_t = \rho h_{t-1} + u, \qquad 0 \leq \rho < 1, \tag{115}$$

with $h_0 = 0$ and constant input u. Then

$$h_t = \sum_{k=0}^{t-1} \rho^k u = \frac{1 - \rho^t}{1 - \rho} u, \tag{116}$$

and hence

$$h_\infty = \frac{1}{1 - \rho} u. \tag{117}$$

Thus, when ρ is close to one, persistent writes can accumulate to a large magnitude despite asymptotic stability.

Mamba-3 introduces an additional phase degree of freedom. Consider the corresponding rotating recurrence

$$z_t = \rho e^{i\theta} z_{t-1} + u, \qquad \rho < 1. \tag{118}$$

For $z_0 = 0$,

$$z_t = \frac{1 - (\rho e^{i\theta})^t}{1 - \rho e^{i\theta}} u, \tag{119}$$

so that

$$\boxed{\|z_\infty\| = \frac{\|u\|}{|1 - \rho e^{i\theta}|}}. \tag{120}$$

Without rotation, $\theta = 0$, this reduces to $\|u\|/(1 - \rho)$, which can become large when $\rho \approx 1$. With nonzero phase, however, the accumulation is controlled by the complex gap $|1 - \rho e^{i\theta}|$ rather than the

radial gap $1-\rho$ alone. Consequently, a near-unit retention modulus does not necessarily imply large state growth: rotation can modulate constructive accumulation without changing the radial retention factor.

Geometrically, purely radial dynamics repeatedly accumulate collinear contributions,

$$u,\ \rho u,\ \rho^2 u, \dots,$$

whereas rotating dynamics accumulate

$$u,\ \rho e^{i\theta} u,\ \rho^2 e^{i2\theta} u, \dots,$$

whose directions vary over time.

C.4 Householder Reflections for Rotation

For a unit-norm key k, setting $\beta = 2$ gives the Householder reflection $H_k = I - 2kk^\top$. In the two-dimensional CKDA construction, let

$$k = (\cos(\theta/2), \sin(\theta/2))^\top, \qquad D_\alpha = \mathrm{Diag}(\alpha, 1), \qquad -1 < \alpha < 0.$$

For the chosen key k, the reflection matrix is

$$\begin{aligned} H_k &= I - 2kk^\top \\ &= \begin{bmatrix} 1 - 2\cos^2(\theta/2) & -2\cos(\theta/2)\sin(\theta/2) \\ -2\cos(\theta/2)\sin(\theta/2) & 1 - 2\sin^2(\theta/2) \end{bmatrix} \\ &= \begin{bmatrix} -\cos\theta & -\sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}. \end{aligned} \tag{121}$$

Since $\alpha < 0$, the diagonal gate decomposes as

$$D_\alpha = H_{e_1} \mathrm{Diag}(-\alpha, 1), \qquad H_{e_1} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{122}$$

The two reflections compose to a rotation:

$$H_k H_{e_1} = \begin{bmatrix} -\cos\theta & -\sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} = R(\theta). \tag{123}$$

Therefore,

$$T_{\alpha,\theta} = H_k D_\alpha = R(\theta)\,\mathrm{Diag}(-\alpha, 1). \tag{124}$$

D Implementation Details

D.1 State Tracking Tasks

Following the setup of Lahoti et al. (2026), we use a sequence-length curriculum in which the minimum sequence length is fixed at 3, while the maximum sequence length increases linearly from 40 to 160. For Parity and Mod-3, this increase spans the first 6,000 optimization steps, after which the maximum remains fixed at 160; for modular arithmetic, it spans all 10,000 steps. For each task, training lasts for 10^4 optimization steps with a batch size of 256. We use one-layer models for Parity and Mod-3, and three-layer models for the modular-arithmetic tasks. The state dimension is fixed at 64. We optimize all models with AdamW, fix $d_{\text{model}} = 64$, and sweep eight learning rates logarithmically spaced between 10^{-4} and 10^{-2}. We select the hyperparameter configuration with the highest validation accuracy and report its performance at sequence length 224.

To investigate the effect of damping regime on state-tracking performance, we construct three variants based on our second-order reparameterization in Eq. 18. Specifically, we set $\kappa_t = -\nu_t^2$, $\kappa_t = 0$, and $\kappa_t = r_t^2$ for the underdamped, critically damped, and overdamped variants, respectively. For the overdamped variant, we parameterize $r_t = c a_t \tanh(\nu_t)$ with $0 < c \leq 1$, ensuring $0 \leq \kappa_t < a_t^2$. These constraints enforce the corresponding spectral families throughout training. The remaining architecture and task-specific training protocol are identical across the three variants.

D.2 Oscillation across Language Tasks

We evaluate official pretrained Mamba-3 1.5B SISO and MIMO models[1] on factual relations from the Relations dataset (Hernandez et al., 2024) and on LAMBADA (Paperno et al., 2016). Following the use of layer windows for interventions in Mamba (Sharma et al., 2024), we apply phase removal within contiguous layer windows, sweeping both the window size and its position across model depth. Within each selected window, we set the rotation angles to zero. We evaluate the same examples with and without the intervention, reporting the accuracy difference in percentage points.

D.3 From-Scratch Pretraining

We train Mamba-3 MIMO 187M and 444M models from scratch on 100B tokens of the FineWeb-Edu dataset (Penedo et al., 2024) with the Llama-3.1 tokenizer (Grattafiori et al., 2024) at a 2K context length. The complete training configuration is summarized in Table 10.

To assess length extrapolation, we evaluate our pretrained models on LongBench-E (Bai et al., 2024) and PG-19 (Rae et al., 2019). For PG-19, we follow prior window-based evaluations (Ben-Kish et al., 2025; Ye et al., 2025). For each context length $C \in \{1, 2, 4, 8, 16, 32, 64\} \times 10^3$ tokens, we include all books containing $L > C$ tokens. From each book, we extract length-C windows beginning at positions $k\Delta$, where $k = 0, 1, \ldots, k\Delta < L-C$, and $\Delta = \max(10, \lfloor (L-C)/10 \rfloor)$. We reset the recurrent state at the start of each window and score only its final 100 tokens, conditioned on the preceding tokens in that window. We report perplexity as the exponential of the mean negative log-likelihood over all scored tokens.

To test whether oscillatory dynamics improve performance beyond the 2K training context across datasets, we evaluate our pretrained models on CodeParrot (Tunstall et al., 2022), MATH-Hard (Hendrycks et al., 2021), TriviaQA (Joshi et al., 2017), and SlimPajama (Soboleva et al., 2023). Following Grazzi et al. (2025), we tokenize each dataset with the pretraining tokenizer, concatenate its documents with EOS separators, and split the resulting stream into 4096-token sequences. We reset the recurrent state at the start of each sequence and carry it through four consecutive 1024-token position bins. For each bin, we exponentiate the mean negative log-likelihood of its tokens across sequences to obtain perplexity. The first two bins cover positions within the 2048-token training context, while the last two evaluate positions beyond it.

To test the zero-shot language modeling capabilities, we use lm-evaluation-harness (Gao et al., 2024) to evaluate our pretrained models on LAMBADA (Paperno et al., 2016), HellaSwag (Zellers et al., 2019), PIQA (Bisk et al., 2020), Arc-Easy/Arc-Challenge (Clark et al., 2018), WinoGrande (Sakaguchi et al., 2021) and OpenBookQA (Mihaylov et al., 2018).

D.4 Pretrained Model Phase Intervention

We conduct the evaluations in Sec. 5.3 using official pretrained Mamba-3 1.5B SISO and MIMO checkpoints.

Learned Phase Dynamics. To visualize phase dynamics during natural-language processing, we analyze the first 16,000 tokens of PG-19 validation books and record the angle projections and time steps. We reconstruct the cumulative phase using the official phase-accumulation kernel at its native numerical precision. For each model, we examine all phase channels and select trajectories near the 10th, 50th, and 99th percentiles of phase travel among channels that complete at least one cycle. These trajectories illustrate a range of slow to fast rotary modulation on natural-language inputs.

Interventions on Phase Structure. We apply four interventions using the PG-19 test protocol described above. Each intervention acts across all layers and heads throughout the sampled window. At each context length, native and intervened runs use the same books, window positions, and scored tokens.

Decay permutation reassigns decay coefficients across heads within each layer while preserving each head's time step:

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

[1] https://huggingface.co/state-spaces

Table 10: Implementation details for from-scratch pretraining on FineWeb-Edu 100B tokens.

parameters	**187M**	**444M**
Model Architecture		
Layers	12	24
d_{model}	768	1024
d_{state}	128	128
Expansion Factor	2	2
Head Dimension	64	64
Intermediate Dimension	1280	1792
Number of Groups	1	1
MIMO Rank	4	4
Chunk Size	16	16
Rotary Channel Fraction	$0/0.5$	$0/0.5$
Optimization		
Optimizer	AdamW	AdamW
Peak LR	3×10^{-3}	1.5×10^{-3}
Minimum LR	10^{-5}	10^{-5}
LR Schedule	Cosine	Cosine
Warmup Steps Percentage	0.375%	0.375%
β_1, β_2	0.9, 0.95	0.9, 0.95
Weight Decay	0.1	0.1
Gradient Clip	1.0	1.0
Global Batch Size	256	256
Precision	BF16 (FP32 params/ optimizer states)	BF16 (FP32 params/ optimizer states)

where π_ℓ is a fixed cyclic permutation of the heads in layer ℓ. Phase dynamics remain unchanged.

Phase shuffling tests whether phase increments must remain assigned to their original rotary pairs. At each layer, head, and token, we permute all rotary pair assignments before phase accumulation:

$$\widetilde{\delta\phi}_{\ell,h,j,t} = \delta\phi_{\ell,h,\sigma_{\ell,h,t}(j),t}.$$

Thus, pair j receives the increment originally computed for pair $\sigma_{\ell,h,t}(j)$. The permutation preserves the set of phase increments within each head and token but changes which pair accumulates each increment. Time steps and radial decay remain unchanged.

Phase removal sets the raw angle projections to zero, so all phase increments vanish and the resulting rotations are identities.

Phase reversal negates each phase increment before accumulation, reversing the direction of rotation.

We report token-aggregated perplexity and the ratio $\mathrm{PPL}_{\text{intervened}}/\mathrm{PPL}_{\text{native}}$ at each context length.

Phase and Recurrent State Accumulation. We compare native dynamics and phase removal on PG-19 validation books with at least 64,000 tokens. For each book and condition, we process the first 64,000 tokens from a zero recurrent state using teacher forcing and the recurrent decoding kernel. Measurements at 16K and 64K use the same uninterrupted trajectory.

In the kernel's state coordinates, we write

$$S_t = P_t + U_t, \qquad P_t = \alpha_t S_{t-1},$$

where U_t is the input contribution. We measure propagation–write alignment as

$$c_t = \frac{\langle P_t, U_t \rangle_F}{\|P_t\|_F \|U_t\|_F + 10^{-12}},$$

using all state coordinates. We average c_t over zero-indexed positions $[2000, L)$, take the median across the 24×64 layer–head pairs, and average equally across books. The RMS state norm is

$$\mathrm{RMS}(L) = \frac{1}{N} \sum_{b=1}^{N} \sqrt{\mathrm{median}_{\ell,h} \left[\frac{1}{L - 2000} \sum_{t=2000}^{L-1} \left\| S_{\ell,h,t}^{(b)} \right\|_F^2 \right]}.$$

Here, N is the number of books, L is the evaluated context length, and $S_{\ell,h,t}^{(b)}$ denotes the recurrent state of layer ℓ and head h at token position t in book b. The Frobenius norm includes all state coordinates without dimension normalization. Both statistics exclude the first 2,000 tokens to reduce initialization transients.

Phase Consistency Between Writing and Readout. We selectively remove the rotary transformation from the write vector B, the readout vector C, or both, only during the final 100 prediction steps of each 64,000-token window. The preceding context is processed with native phase dynamics, and the recurrent state is carried into the intervention region without resetting. Removing phase replaces the corresponding rotation with the identity while retaining the write and readout vectors. The intervention applies to all layers, heads, and rotary pairs, including all ranks in MIMO. We evaluate all four conditions on all PG-19 test books, scoring only the final 100 target tokens.

D.5 LongBench-E Task Abbreviations

For compact presentation, we use the following abbreviations for LongBench-E tasks throughout the paper.

Table 11: Task-name abbreviations used for LongBench-E evaluation.

Abbr.	Task	Abbr.	Task
2WM	2WikiMQA	GR	GovReport
HQA	HotpotQA	LCC	Long Code Completion
MQA	MultiFieldQA-en	MN	MultiNews
PC	Passage Count	PR	PassageRetrieval-en
QA	Qasper	RB	RepoBench-P
SS	SAMSum	TR	TREC
TQA	TriviaQA		

E Additional Results

E.1 Phase Structure Induced by Oscillation

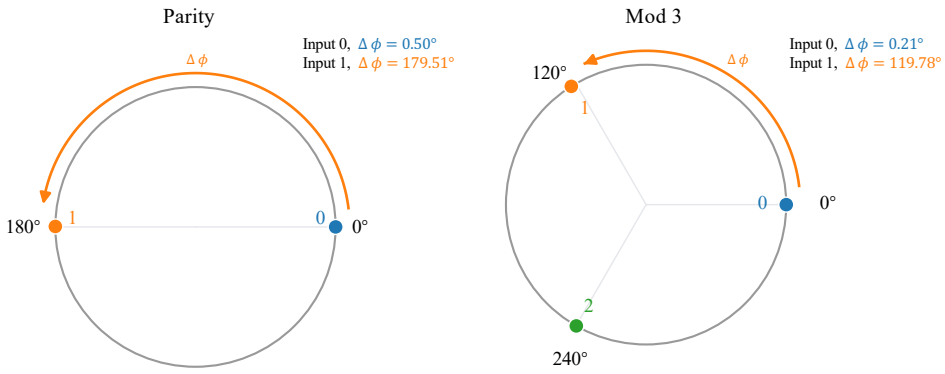

Figure 5: Task-aligned phase structure in selected two-dimensional state pairs. Input 0 induces little phase change, whereas input 1 induces a rotation close to $180°$ for parity and $120°$ for mod-3.

Figure 5 illustrates task-aligned phase structure in selected two-dimensional neural state pairs. For parity, one pair changes phase by $0.50°$ on input 0 and $179.51°$ on input 1. For mod-3, a selected pair shows corresponding changes of $0.21°$ and $119.78°$. These angles are close to the increments of an ideal phase code for counting ones modulo m: input 0 leaves the phase unchanged, while input 1 advances it by $360°/m$. Repeated updates would then place the two parity classes near $0°$ and $180°$, or the three mod-3 classes near $0°$, $120°$, and $240°$.

This interpretation gives a concrete role to input-dependent oscillation in state tracking. In a direct rotary implementation, the phase increment must depend on the current symbol: a fixed rotation on every step would also advance the phase on input 0 and would not implement this particular count-of-ones code. The observed pairs are therefore consistent with the model using conditional rotations to represent modular state, although they do not establish that every state pair follows the same geometry.

E.2 Length Generalization on Other Datasets

As shown in Fig. 6, oscillatory models generally achieve lower perplexity, with more consistent improvements at the larger model scale. At 187M parameters, the oscillatory model has lower perplexity on CodeParrot, TriviaQA, and SlimPajama, but higher perplexity on MATH-Hard. At 444M parameters, it has lower perplexity on all four datasets across the evaluated context lengths. The differences beyond the 2K training context are most apparent on CodeParrot and TriviaQA, while those on SlimPajama remain small. These results provide additional evidence of improved performance at evaluated lengths up to 4K, with benefits that vary across datasets and model scales.

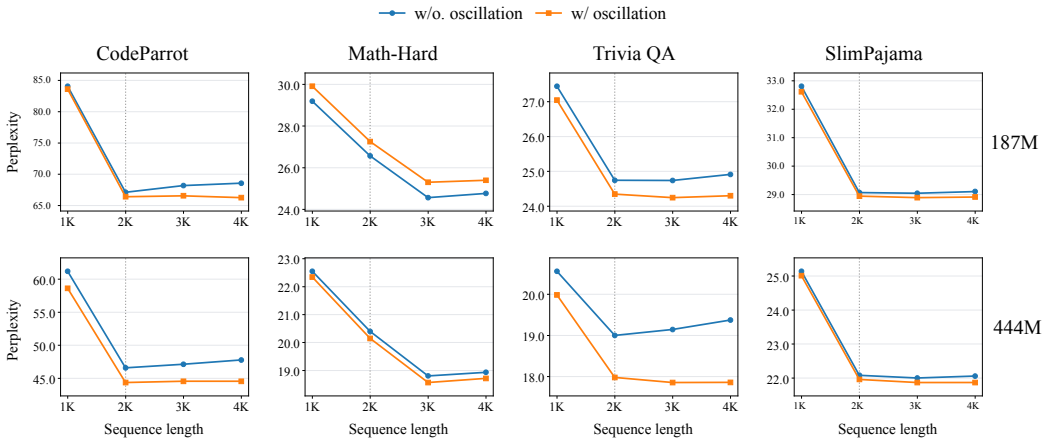

Figure 6: Perplexity across context lengths on CodeParrot, MATH-Hard, TriviaQA, and SlimPajama for models with and without oscillatory dynamics. Rows correspond to the 187M (top) and 444M (bottom) models. Dotted vertical lines mark the 2K training context length.

E.3 Training Dynamics

We further examine how the advantage of oscillatory dynamics develops throughout pretraining. Figure 7 shows the validation perplexity trajectories of the 187M and 444M models. At both model scales, the oscillatory variant reaches lower validation perplexity toward the end of training.

Figures 8 and 9 track PG-19 perplexity at different training checkpoints across context lengths from 2K to 64K. At the end of training, the oscillatory variant achieves lower perplexity at every evaluated context length for both model sizes. The separation is more pronounced at longer context lengths, particularly for the larger 444M model: while the two variants remain close at short contexts, their perplexity gap becomes substantially larger at 16K–64K. These trajectories suggest that the benefit of oscillatory dynamics becomes more pronounced for larger models operating over long contexts.

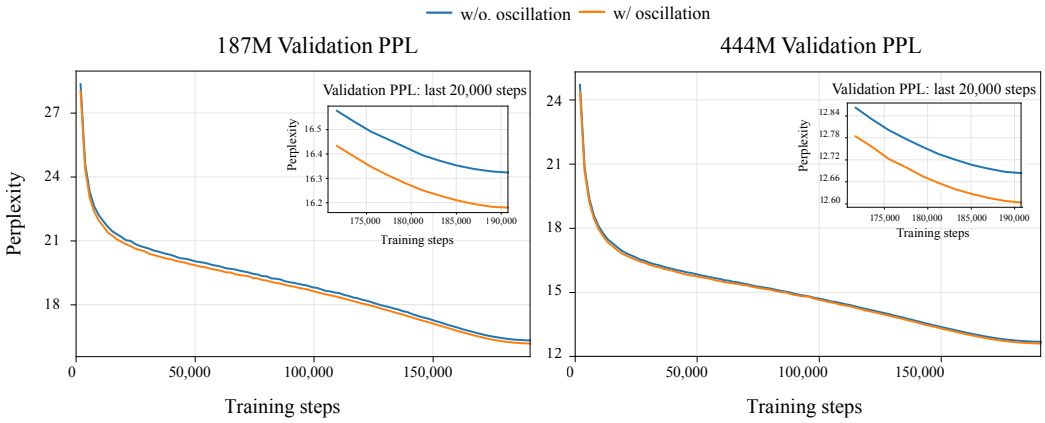

Figure 7: Validation perplexity throughout pretraining. Training trajectories of Mamba-3 187M and 444M with and without oscillatory dynamics. Insets show the final 20,000 training steps.

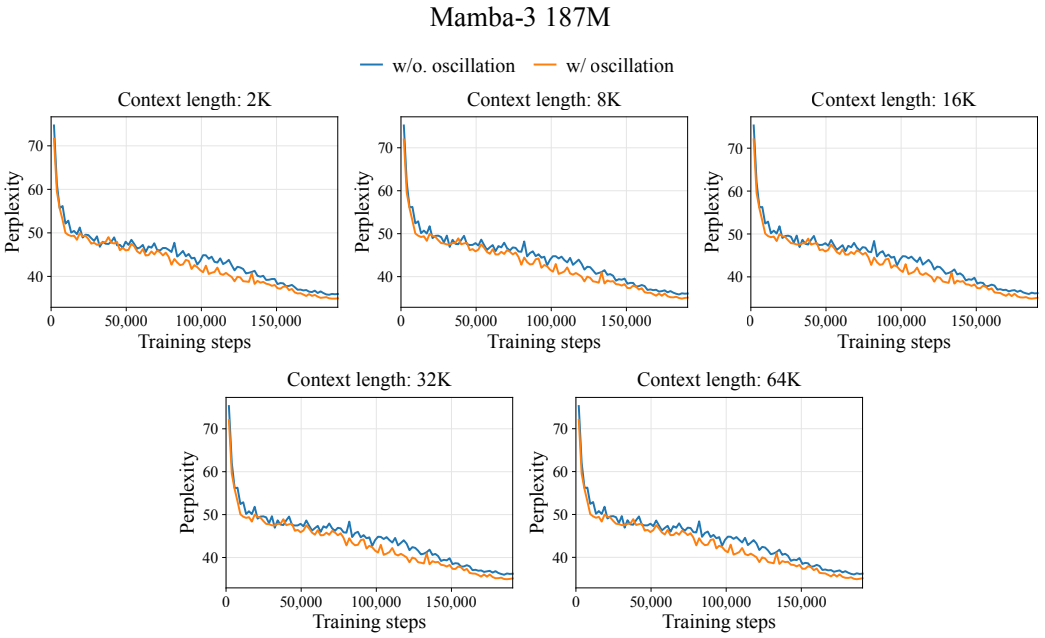

Figure 8: PG-19 perplexity throughout pretraining for Mamba-3 187M. PG-19 perplexity is evaluated at context lengths from 2K to 64K at different training checkpoints. The oscillatory variant achieves lower perplexity across all evaluated context lengths toward the end of training.

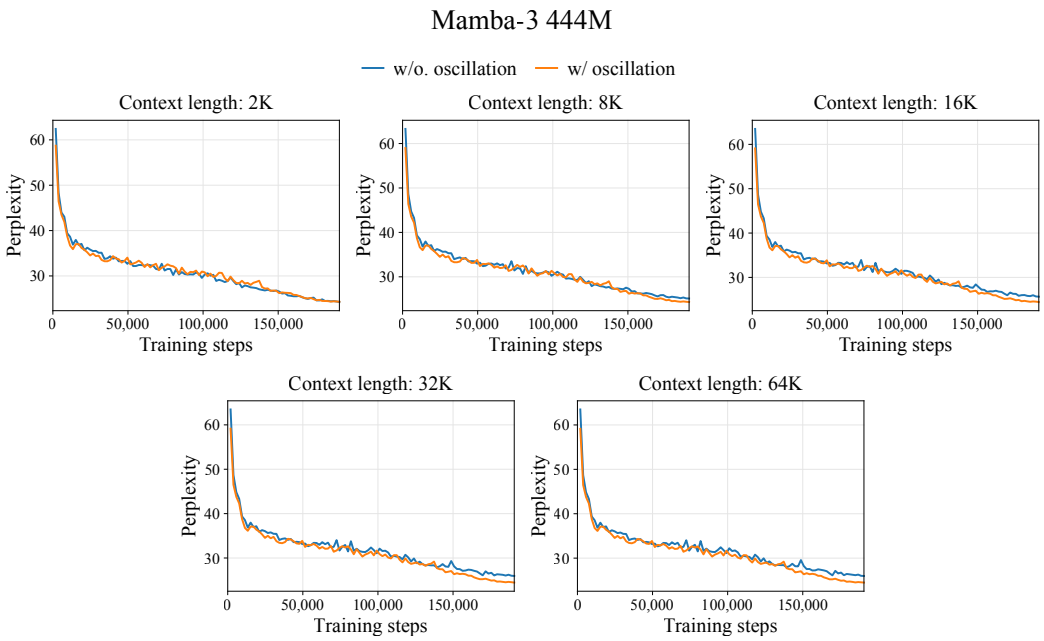

Figure 9: PG-19 perplexity throughout pretraining for Mamba-3 444M. PG-19 perplexity is evaluated at context lengths from 2K to 64K at different training checkpoints. The oscillatory variant achieves lower perplexity across all evaluated context lengths toward the end of training, with a more pronounced separation at longer contexts.