Lyapunov Spectral Analysis of Loop Transformer Dynamics

Anonymous Authors[1]

Abstract

Loop Transformers iterate a shared block of layers, defining a discrete dynamical system over hidden states. Existing characterizations rely on attention or hidden-state similarity, which cannot distinguish slow convergence, marginal stability, and chaos. We compute the Lyapunov spectra of two loop transformers and find a dichotomy in dynamics: while Ouro-1.4B is mildly chaotic and rules out convergence under the measured finite-time dynamics, Huginn-0125 converges uniformly in all dimensions. A per-sublayer attribution provides a mechanistic account of how each regime is produced. Both architectures exhibit near-cancellation between large opposing contributions of different layers, however the patterns differ significantly. Ouro distributes compression and expansion across 25 sublayers, with direction-selective late layers and direction-blind RMSNorm jointly producing a wide spectrum. Huginn concentrates the entire cancellation between the input-injection adapter and the first core block. This supports the empirical observation that input injection encourages fixed-point convergence hinges on an architectural balance between two blocks. A measurement of the first Lyapunov exponent across 8 Huginn training checkpoints further shows the regime emerges early and remains stable. Ultimately, we establish Lyapunov spectra as a rigorous lens for characterizing the stability regimes and mechanistic behavior of loop transformers.

1. Introduction

Loop transformers iterate a shared block of layers T times, producing a discrete dynamical system $z_{t+1} = f_\theta(z_t)$ over hidden states (Geiping et al., 2026; Zhu et al., 2025). Because the same f_θ is applied at every step, the recurrence defines a proper iterated map and admits dynamical systems analysis that is not available for standard feedforward networks.

Loop transformers are designed to scale test-time compute by increasing recurrence depth (Geiping et al., 2026; Zhu et al., 2025), often well beyond what they were trained on. Whether this additional depth produces useful computation depends entirely on the long-term behavior of the iterated map: convergence, marginal stability, and chaos have very different implications for what the recurrence is doing. Existing approaches are either theoretical, analyzing simplified components in isolation (Dong et al., 2021; Wu et al., 2024), or empirical but qualitative (Blayney et al., 2026).

Blayney et al. (2026) recently characterized loop transformer dynamics through attention pattern similarity and hidden state cosine similarity, observing a dichotomy: Huginn-0125 converges to a fixed point, while Ouro-1.4B exhibits a wandering trajectory that approaches but never reaches a fixed point. Their analysis cannot distinguish three qualitatively different dynamics: slow convergence, marginal stability, or chaos. These have very different implications for the underlying computational mechanism. Several prior works (Bansal et al., 2022; Anil et al., 2022; Blayney et al., 2026) additionally observe empirically that input injection encourages fixed-point convergence in recurrent loop transformer architectures, but the architectural components that determine the regime have not yet been identified.

We address both questions using Lyapunov spectral analysis: which regime does a given loop transformer exhibit, and which architectural components produce it?

The Lyapunov spectrum is a classical tool from dynamical systems that characterizes the asymptotic stability of an iterated map by measuring the exponential rate at which nearby trajectories separate or converge. Applied to loop transformers, it resolves the ambiguity in non-convergence, and the per-sublayer decomposition reveals which architectural components are responsible for that regime.

Contributions. (i) We introduce Lyapunov spectral analysis as a framework to study loop transformer dynamics. Computing the top-5 spectrum for both Ouro-1.4B and

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Huginn-0125 distinguishes a chaotic regime from a contractive one (§3.1).

(ii) We apply this framework to analyze per-sublayer attribution by re-orthogonalizing after each sublayer instead of each loop. This provides a quantitative mechanism for the empirical observation that input injection encourages fixed-point convergence.

(iii) We apply the same framework to study training dynamics, measuring λ_1 across 8 publicly available Huginn training checkpoints and finding that the dynamical regime is established early and remains stable thereafter (§3.3).

2. Background

2.1. Loop Transformers

We analyze Ouro-1.4B (Zhu et al., 2025) and Huginn-0125 (Geiping et al., 2026). Ouro applies 24 transformer blocks $B_l(X) = X + \mathrm{Attn}(X) + \mathrm{MLP}(X + \mathrm{Attn}(X))$ followed by RMSNorm: $f(X) = \mathrm{RMSNorm}(B_{24}(\cdots B_1(X)))$. We refer to each transformer block B_l as *layer* and each application of the iterated map $f(X)$ as *loop*. It is trained with $T{=}4$ recurrences and a learned early-exit gate, which we bypass to study long-term dynamics. Huginn iterates a 4-layer core block with mean training recurrence $T{=}32$ and uses input injection: a projection of the prelude output is added to the recurrent state at each step. Unlike Ouro, Huginn is designed for any number of inference iterations.

2.2. Lyapunov Exponents

The Lyapunov exponents characterize the exponential rate at which nearby trajectories separate, $|\delta_t| \approx e^{\lambda t}\,|\delta_0|$.

Consider a discrete-time map $f : \mathbb{R}^d \to \mathbb{R}^d$ and it's induced trajectory from an initial state z_0 given by $z_{n+1} = f(z_n)$. An infinitesimal perturbation $\delta_0 \in \mathbb{R}^d$ propagates under linearized dynamics $\delta_{n+1} = J_n \delta_n$ where $J_n = \partial f/\partial z\big|_{z_n}$ is the Jacobian evaluated along the trajectory. After n step, we have

$$\delta_n \;=\; \tilde{J}_n\,\delta_0, \qquad \tilde{J}_n \;\equiv\; J_{n-1}J_{n-2}\cdots J_0.$$

The *Lyapunov exponent* associated with the direction δ_0 is the asymptotic exponential rate of separation,

$$\lambda(z_0, \delta_0) \;=\; \lim_{n\to\infty} \frac{1}{n} \ln \frac{\|\tilde{J}_n \delta_0\|}{\|\delta_0\|}.$$

The Lyapunov exponents associated with all d orthogonal directions form the Lyapunov spectrum. We adopt the convention $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$ indexed from one.

The maximal exponent λ_1 determines the long-term dynamics of the trajectory. A positive $\lambda_1 > 0$ is the defining feature of chaos, while $\lambda_1 < 0$ indicates contraction to a fixed point. With $\lambda_1 \approx 0$ the trajectory is marginally stable.

The Lyapunov exponents are given equivalently by the singular values of the Jacobian product,

$$\lambda_i \;=\; \lim_{n\to\infty} \frac{1}{n} \ln \sigma_i(\tilde{J}_n).$$

We compute the top-k exponents using the Benettin algorithm (Benettin et al., 1980), which periodically reorthogonalizes a set of probe vectors to prevent collapse onto the dominant expanding direction (Appendix A).

3. Spectral Analysis

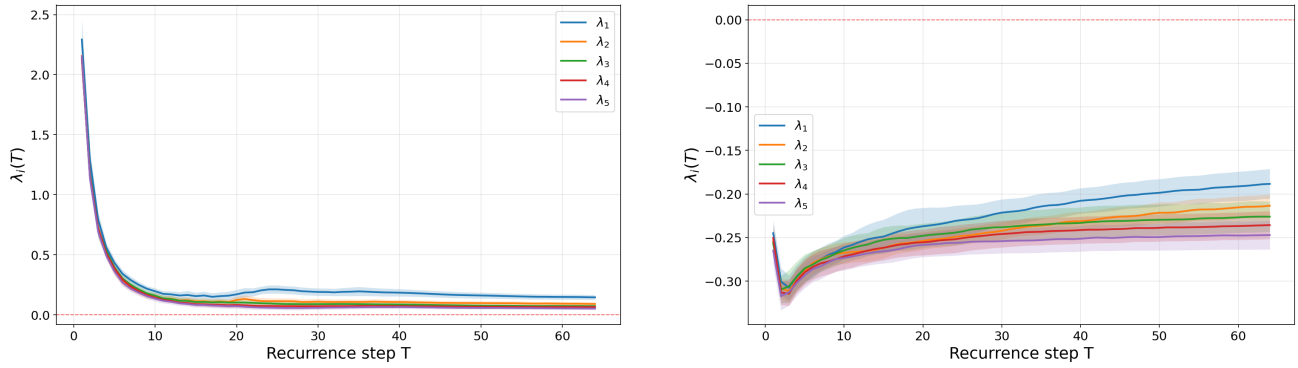
3.1. Loop Convergence

We compute the top $k{=}50$ Lyapunov exponents for both models using $T{=}64$ recurrence steps across 20 prompts (Appendix B). To analyze long-term dynamics, we deliberately choose to extend beyond Ouro's training limit of $T{=}4$. Calculating the entire Lyapunov spectrum of the models is computationally infeasible, which is why we restrict our analysis to the only the top-50 exponents. We find the two models exhibit qualitatively different regimes matching the observations of Blayney et al. (2026). For Ouro, the top 24 exponents are positive or near zero, indicating chaotic dynamics. Huginn's top 50 exponents lie entirely in $[-0.33, -0.21]$, indicating uniform contraction across all measured directions. Fig. 1 shows the top-5 exponents for both models, plotted as running averages over recurrence steps and averaged across all prompts.

The two Lyapunov spectra quantitatively explain the dynamics observed by Blayney et al. (2026). Huginn's hidden state contracts exponentially, with $\|z_{t+1} - z_t\|$ decaying from ~ 100 to $\sim 10^{-4}$ over 64 steps and reaching its fixed point by step $\sim$32. This is consistent with the negative spectrum.

For Ouro, the positive exponents establish an even stronger claim than non-convergence: the trajectories are not slowly approaching a fixed point and will not do so for any number of iterations. Blayney et al. (2026) described these trajectories as "approximately constant trajectory". Our Lyapunov analysis strengthens this interpretation: because $\lambda_1 > 0$, the observed near-cyclic behavior is unlikely to correspond to a stable periodic orbit, and is more naturally explained as transient motion near unstable periodic structure within a strange attractor.

This dichotomy is also visible in the residual-stream cosine similarity as shown in Fig. 4. Ouro's contracting directions cause corresponding layers across loops to grow more similar (brightening off-diagonal blocks) but the expanding directions prevent saturation; Huginn saturates near 1.0 across most layer pairs.



(a) Ouro-1.4B *(b)* Huginn-0125

Figure 1. Top-5 Lyapunov exponent convergence over 64 recurrence steps averaged over 20 prompts. Ouro: top-24 positive exponents. Huginn: all measured exponents negative.

3.2. Intra-Loop Attribution

We decompose each Lyapunov exponent into per-sublayer contributions by performing QR re-orthogonalization after each sublayer rather than once per loop step. We compute the top-1 Lyapunov exponent for both models using $T{=}32$ recurrence steps across the same 20 prompts (Appendix B). For Ouro this gives 25 sublayers (24 decoder layers plus the final RMSNorm) and 5 sublayers for Huginn (the input-injection adapter plus the 4 core block layers). We verify the per-sublayer log-growth factors sum exactly to the full-step Lyapunov exponent.

Both architectures exhibit near-cancellation between large positive and negative per-sublayer contributions to the overall Lyapunov Exponent. However, the per-layer contribution breaks down differently. Fig. 2 compares the contributions of each layer to λ_1 for both models.

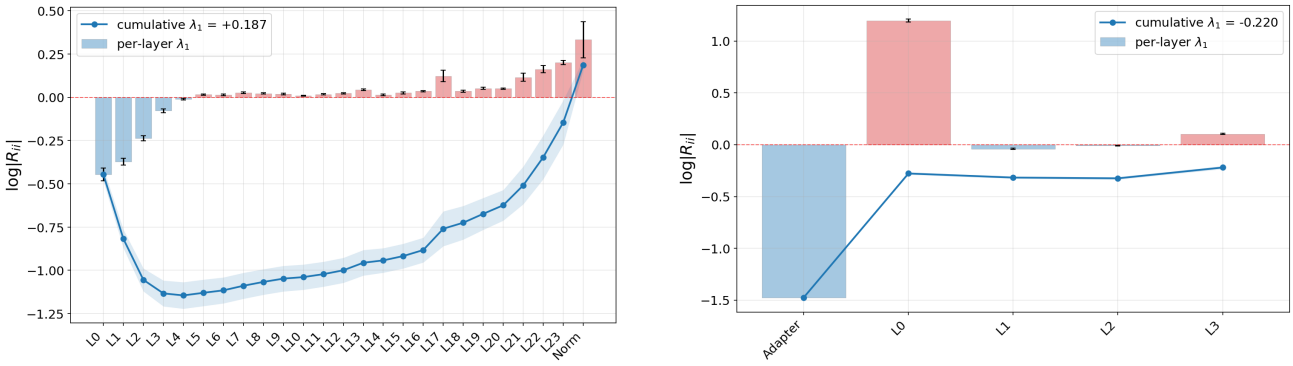
Ouro. The per-sublayer contribution increases monotonically through the loop: L0 is heavily contracting with a negative exponent, rising steadily to L23 with the final norm being most expansive with the highest positive exponent. The heatmap (Fig. 5a) reveals that late layers and the norm expand through different mechanisms: late layers are *direction-selective*, preferentially amplifying the top Lyapunov directions, while the final norm is *direction-blind*, contributing roughly uniformly across all measured directions. Early layers reduce state magnitude, RMSNorm's projection back onto the hypersphere amplifies all directions roughly equally. The late layers create the directional structure that distinguishes expanding from contracting directions. Tracking λ_1 through the layers (Fig. 2a) reveals how the exponent remains net positive. λ_1 drops to a minimum at L4, recovers slowly and is only pushed above zero only by the norm.

Huginn. Even with only 4 layers, Huginn exhibits a similar disparity between positive and negative exponents across its loop pass. The pattern, however, differs from Ouro significantly. The input-injection adapter contracts strongly, the first core block immediately re-expands, and L1–L3 are essentially fine-tuning with Lyapunov exponents close to zero. Two large primitives in opposition determine the regime, while the remaining sublayers contribute negligibly. Unlike Ouro, both primitives are direction-blind. The heatmap rows are nearly uniform across the 50 measured directions, and the top-5 exponents differ only at the third decimal. This explains the narrow negative Huginn spectrum band of the top-50 Lyapunov exponents (-0.21 to -0.33) compared to Ouro's wider range in the top-50 ($+0.17$ to -0.04): without direction-selective Jacobians, there is no mechanism to create the directional structure that produces a wide spectrum.

3.3. Early-Training Dynamics

The previous analysis suggests that dynamical regime is determined by architectural choices. Under this hypothesis, the differences in Lyapunov exponents should be visible from early in training rather than emerging gradually. We test this on Huginn by measuring λ_1 across the 8 publicly available intermediate training checkpoints plus the final model, using a single-probe Benettin estimator at $T{=}32$.

Across all 9 measured points λ_1 lies in a narrow band $[-0.27, -0.22]$ (Fig. 3). The across-checkpoint standard deviation (0.015) is less than half the per-prompt measurement noise (0.032). Huginn is contractive throughout all checkpoints and we observe no sign-crossing nor any monotone trend. This is consistent with the per-layer finding. Since the contractive dynamics emerge early and remain stable, it might suggest that it is caused by architecture like the structural balance between adapter contraction and core block expansion. We caveat that the first 13% of training is not covered by public checkpoints, leaving open whether



(a) Ouro-1.4B *(b)* Huginn-0125

Figure 2. Per-sublayer contribution to λ_1. Both architectures exhibit near-cancellation between large opposing contributions. Ouro: early decoder layers contract and late layers plus the final norm expand, with λ_1 slightly above zero. Huginn: the input-injection adapter contracts and the first core block immediately re-expands, with λ_1 slightly below zero.

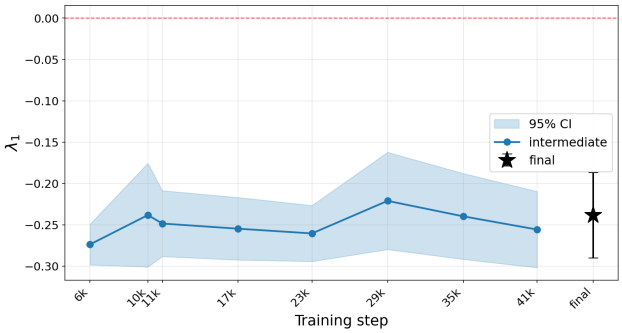


Figure 3. λ_1 across Huginn-0125 training checkpoints averaged over 5 prompts. Black star: final `main` checkpoint. Across all 9 measured points λ_1 lies in a narrow, negative band with no sign-crossing or monotone trend.

a transient regime exists at initialization that resolves before the first observed checkpoint. Equivalent intermediate checkpoints are not publicly available for Ouro, so we cannot make the analogous comparison there.

3.4. Impact of Architectural Choices

The Huginn decomposition gives a quantitative mechanism for an empirical observation made by Bansal et al. (2022), Anil et al. (2022), and Blayney et al. (2026): input injection encourages fixed-point convergence. The adapter contributes -1.47 per step against the core block's $+1.17$. The regime hinges on this balance. Removing or weakening input injection should cause expansion and likely flip Huginn into a chaotic regime. We leave this ablation for future work.

4. Conclusion

We provide the first Lyapunov spectral analysis of loop transformers. The spectrum distinguishes two dynamical regimes: Ouro-1.4B is mildly chaotic and cannot converge to a fixed point or periodic orbit, while Huginn-0125 contracts uniformly. Per-sublayer decomposition reveals that both architectures produce their regimes through near-cancellation between large opposing contributions, but the cancellation is implemented by different primitives. Ouro distributes compression and expansion across 25 sublayers, with direction-selective late layers and direction-blind RMSNorm jointly producing a wide spectrum. Huginn concentrates the entire cancellation between the input-injection adapter and the first core block. The fixed-point convergence induced by input injection hinges on the balance between adapter contraction and core block expansion.

Limitations and future work. We analyze 20 prompts, 50 of the $\sim$16,000 exponents, and two model architectures. The chaotic-vs-contractive classification depends on a fine balance between large opposing contributions, which may not generalize across model scales. Our training-checkpoint measurement shows that λ_1 is stable across the visible portion of Huginn's training, but cannot distinguish whether the regime is set by the architecture itself or established within the unobserved first 13% of training. For Ouro, equivalent intermediate checkpoints are not publicly available. Confirming λ_1 as a general training diagnostic requires denser sampling on self-trained models. Promising further directions include the input-injection ablation predicted by Section 3.2. This would require training matched models with and without input injection to test whether the architectural balance alone determines the regime. Additionally, the positive spectrum could be connected to information production via the Pesin entropy formula by inspecting ergodicity assumptions. Furthermore the analysis can be extended to

other architectures.

References

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *Advances in Neural Information Processing Systems*, 38:41340–41391, 2026.

Rui-Jie Zhu, Zixuan Wang, Kai Hua, Tianyu Zhang, Ziniu Li, Haoran Que, Boyi Wei, Zixin Wen, Fan Yin, He Xing, et al. Scaling latent reasoning via looped language models. *arXiv preprint arXiv:2510.25741*, 2025.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International conference on machine learning*, pages 2793–2803. PMLR, 2021.

Xinyi Wu, Amir Ajorlou, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. On the role of attention masks and layernorm in transformers, 2024a. *URL https://arxiv.org/abs/2405.18781*, page 2, 2024.

Hugh Blayney, Álvaro Arroyo, Johan Obando-Ceron, Pablo Samuel Castro, Aaron Courville, Michael M Bronstein, and Xiaowen Dong. A mechanistic analysis of looped reasoning language models. *arXiv preprint arXiv:2604.11791*, 2026.

Arpit Bansal, Avi Schwarzschild, Eitan Borgnia, Zeyad Emam, Furong Huang, Micah Goldblum, and Tom Goldstein. End-to-end algorithm synthesis with recurrent networks: Extrapolation without overthinking. *Advances in Neural Information Processing Systems*, 35:20232–20242, 2022.

Cem Anil, Ashwini Pokle, Kaiqu Liang, Johannes Treutlein, Yuhuai Wu, Shaojie Bai, J Zico Kolter, and Roger B Grosse. Path independent equilibrium models can better exploit test-time computation. *Advances in Neural Information Processing Systems*, 35:7796–7809, 2022.

Giancarlo Benettin, Luigi Galgani, Antonio Giorgilli, and Jean-Marie Strelcyn. Lyapunov characteristic exponents for smooth dynamical systems and for hamiltonian systems; a method for computing all of them. part 1: Theory. *Meccanica*, 15(1):9–20, 1980.

A. Benettin Algorithm

The naive approach of forming the Jacobian product $\tilde{J}_T = J_{T-1} \cdots J_0$ and taking singular values fails because the product overflows or collapses numerically within a few steps. The Benettin algorithm periodically re-orthogonalizes a set of probe vectors via QR, preventing them from collapsing onto the dominant expanding direction; the diagonals of R accumulate the per-step log-growth factors that converge to the Lyapunov exponents by Oseledets' theorem. We obtain Jacobian-vector products via PyTorch's forward-mode autodiff (JVP).

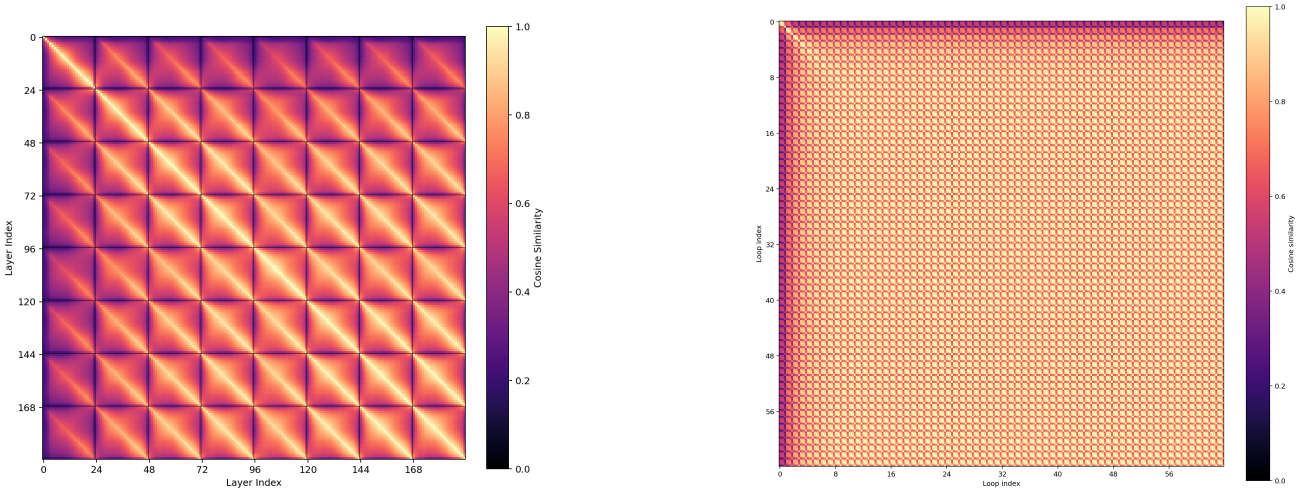
Algorithm 1 Benettin Algorithm (Top-k Lyapunov Exponents)

$Q \leftarrow \mathtt{QR}(\mathtt{randn}(d, k)), \quad S \leftarrow \mathbf{0} \in \mathbb{R}^k$
for $t = 0, \ldots, T-1$ **do**
  $W_{\cdot,i} \leftarrow Df(z_t) \cdot Q_{\cdot,i}$ for $i = 1, \ldots, k$ ▷ via JVP
  $Q, R \leftarrow \mathtt{QR}(W)$
  $S \leftarrow S + \ln|\operatorname{diag}(R)|$
  $z_{t+1} \leftarrow f(z_t)$
end for
return $\lambda_i = S_i/T$

B. Prompts

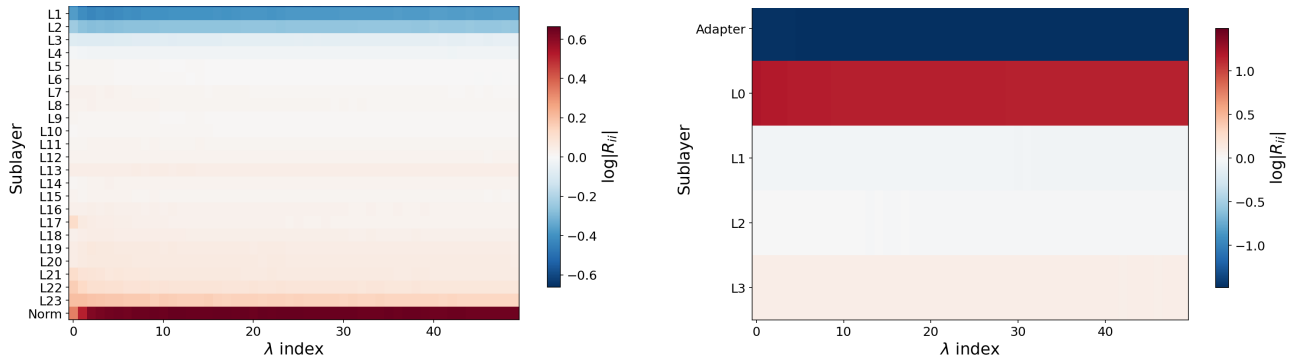
All experiments use 20 prompts spanning declarative, narrative, and instructional registers, not selected based on observed dynamics, such as: "Water is composed of"; "Once upon a time, in a kingdom far away,"; "To bake a cake, first"; "To solve this problem we need to first consider"; "If all birds can fly and a penguin is a bird, then".

C. Additional Plots



(a) Ouro-1.4B *(b)* Huginn-0125

Figure 4. Pairwise cosine similarity of residual stream states across 8 loops. Ouro shows persistent gaps below 1.0. Huginn saturates near 1.0. Different block sizes in the patterns show differing numbers of layers per loop.



(a) Per-sublayer heatmap (Ouro) *(b)* Per-sublayer heatmap (Huginn)