Lyapunov Spectral Analysis of Loop Transformer Dynamics

UNIVERSITY OF CAMBRIDGE

Fynn Kiwitt Christopher Irwin Riccardo Ali Pietro Liò

Department of Computer Science and Technology, University of Cambridge

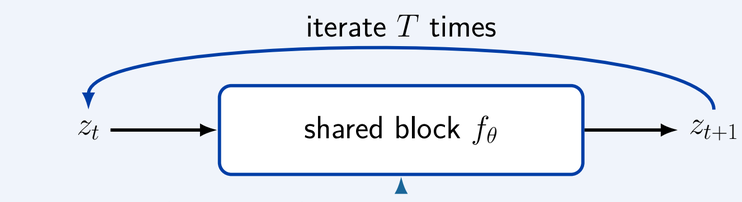
TL;DR: Lyapunov spectra classify loop transformer long-term dynamics. **Ouro-1.4B is mildly chaotic** ($\lambda_1 > 0$ over the measured horizon) while **Huginn-0125 contracts to a fixed point** ($\lambda_1 < 0$).

1 Similarity Can't Tell Convergence from Chaos

- **Loop transformers** iterate a shared block of layers: $z_{t+1} = f_\theta(z_t)$, a discrete dynamical system over hidden states.
- Test-time compute scales with recurrence depth T, often far beyond the training horizon.
- **Contraction** $\rightarrow$ settles to a fixed answer; Extra depth recurrences have a diminishing effect.
- **Chaos** $\rightarrow$ never settles. Small perturbations compound.
- Attention or cosine similarity cannot distinguish slow convergence, marginal stability and chaos.

We ask: which regime does each model exhibit, and which architectural components produce it?

2 Two Loop Transformer Architectures



Simplified schematic

Ouro-1.4B, a looped reasoning LM: 24 decoder blocks + final RMSNorm per loop (d_{model}=2048), trained at T=4, no input injection.

Huginn-0125, a depth-recurrent LM: 4-layer core block per loop (d_{model}=5280), mean training depth $T\approx$32, with input injection.

Prior work observes that input injection encourages fixed-point convergence (Bansal et al. '22, Anil et al. '22, Blayney et al. '26), but it was unknown which components determine the regime.

3 Lyapunov Spectra

A perturbation δ_0 of the hidden state propagates under the linearized dynamics $\delta_{n+1} = J_n\,\delta_n$ with $J_n = \partial f_\theta/\partial z\big|_{z_n}$, giving the exponential separation rate

$$\lambda(z_0, \delta_0) = \lim_{n\to\infty} \frac{1}{n} \ln \frac{\|\tilde{J}_n\,\delta_0\|}{\|\delta_0\|}, \qquad \tilde{J}_n = J_{n-1}\cdots J_0.$$

The singular values of $\tilde{J}_n$ give the full sorted spectrum, $\lambda_i = \lim_{n\to\infty} \frac{1}{n} \ln \sigma_i(\tilde{J}_n)$, and the top exponent classifies the regime:

λ_1<0: **contraction** $\lambda_1\approx$0: marginal λ_1>0: **expansion**

For bounded trajectories, $\lambda_1 > 0$ is standard evidence of chaotic dynamics.

Estimation (Benettin). K probe vectors are re-orthogonalized (QR) after every step. The time-average of $\ln|\mathrm{diag}(R)|$ converges to λ_i (Oseledets' theorem). Jacobian-vector products via forward-mode autodiff. All reported values are finite-time estimates $\lambda_i(T)$ along the model's own trajectory.

4 Two Dynamical Regimes

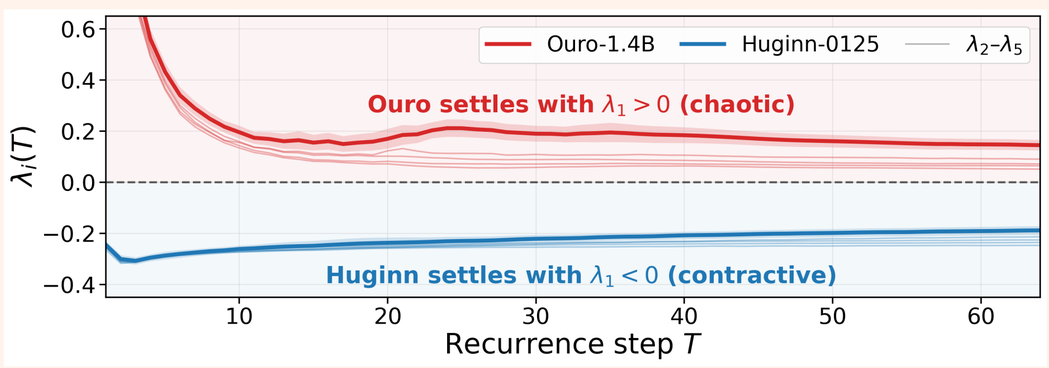


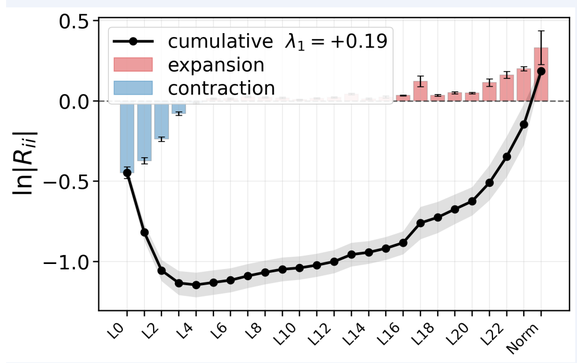
Figure: Running Lyapunov estimates $\lambda_i(T)$ over T=64 recurrence steps (mean ± 95% CI over 20 prompts, K=50 probes). Ouro's λ_1 settles at $+0.145$. Huginn stays below zero throughout.

Ouro: $\lambda_1 = +0.145$
positive on 20/20 prompts (chaotic)

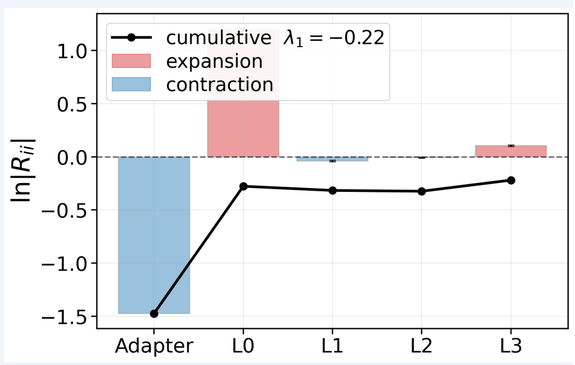
Huginn: $\lambda_1 = -0.21$
all 50 exponents negative (contractive)

- **Ouro** enters chaos regime beyond its training depth (T=4): inside the training window, growth comes almost entirely from the single embedding step (per-pass $\bar{\lambda} = +2.29$), the following passes contract.
- Consistent pass-by-pass expansion appears only from $t \geq 8$.
- $\lambda_1 > 0$ argues against convergence on this horizon. This is consistent with the wandering trajectories of Blayney et al., suggesting motion near an unstable periodic structure rather than a stable cycle.
- Although 30% of single passes still contract, the trajectory expands on average.
- **Huginn** contracts at every depth: $\|z_{t+1} - z_t\|$ decays $\sim$100 $\rightarrow 10^{-4}$, fixed point by step $\approx$32, matching the uniformly negative spectrum.

5 Intra-Loop Attribution



Ouro: 24 layers + Norm



Huginn: Adapter + 4 core layers

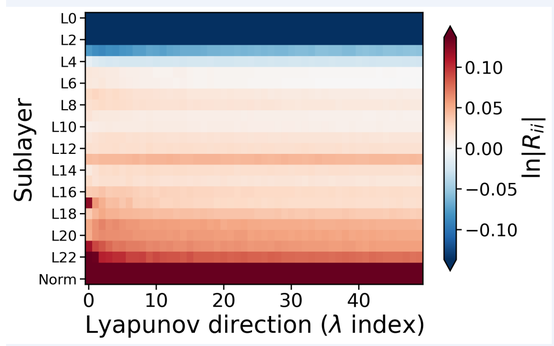
Figure: Per-sublayer contribution to λ_1 (bars) and cumulative sum through one loop (line): QR after every sublayer, so contributions sum exactly to λ_1. T=32, 20 prompts.

Both models balance **large opposing contributions** via different primitives:

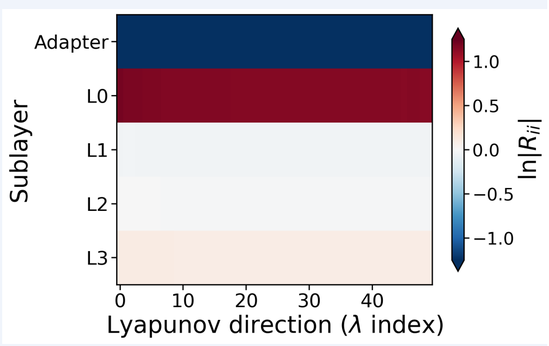
Ouro distributes the λ_1 balance across 25 sublayers: early layers contract, late layers recover, and only the final RMSNorm pushes λ_1 above zero.

Huginn's λ_1 balance is dominated by two components: the input-injection adapter ($\Delta\lambda_1 = -1.47$) and the first core block ($+1.17$). L1–L3 are near-neutral.

6 Differentiating Directions



Ouro: direction-selective



Huginn: direction-blind

Figure: Per-sublayer contribution to each of the 50 tracked orthogonal directions (one per λ_i). Column sums give the spectrum values λ_i. Red = expansion, blue = contraction. Color scales clipped per panel (extreme rows saturate).

Ouro: late layers amplify the top directions preferentially (left edge, L17–L23), creating the spectrum's width ($+0.17$ to -0.04 at T=64). The RMSNorm contribution is nearly uniform across tracked directions, setting its positive offset.

Huginn: both dominant components are nearly uniform across tracked directions, yielding a narrow spectrum $[-0.33, -0.21]$.

7 Consistency over Training

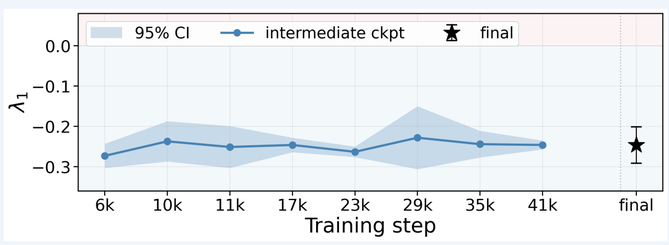


Figure: λ_1 across 8 public Huginn checkpoints + final model (single-probe Benettin, T=32, 5 prompts).

- All nine points in $[-0.27, -0.22]$: no sign change, no trend.
- Consistent with an architectural origin (adapter–core balance). **Caveat:** the first 13% of training is unobserved, so an early transient cannot be ruled out.
- No intermediate checkpoints are public for Ouro, so the analogous comparison is not possible there.
- Cheap to track (K=1 probe): a candidate **training diagnostic** for recurrent architectures.

8 Takeaways

- Lyapunov spectra turn qualitative observations into a **quantitative regime classification**.
- In the two models, regimes emerge from **near-cancellation of large opposing sublayer contributions**. In Huginn, input injection is implicated through the adapter vs. core-block balance.
- **Limitations:** finite-time estimates, top-50 of $\geq$8k exponents, 20 prompts, two models.
- **Next:** run input-injection ablation, explore information-theoretic connection using Pesin entropy and analyze more architectures.

References: Zhu et al. '25 (Ouro), Geiping et al. '26 (Huginn), Blayney et al. '26, Benettin et al. '80, Bansal et al. '22, Anil et al. '22

Correspondence: fk407@cam.ac.uk