Depth-Dynamics Signatures of Conversational Collapse: Finite-Time Lyapunov Analysis of Transformer Forward Passes

Sohail Mohammad
Independent Researcher
sohailmo.ai@gmail.com

February 2026

Abstract

We estimate the top-1 finite-time Lyapunov exponent (λ_1) for transformer depth dynamics using forward-mode automatic differentiation (JVP-based tangent propagation) with QR renormalization, and test whether depth-dynamics summaries predict conversational collapse behavior observed in multi-turn self-play. Across 720 preregistered trajectories (4 conditions $\times$ 36 seeds $\times$ 5 repeats) and 7,200 FTLE computations on three 7B-parameter model families, we find that λ_1 *profile features*—specifically the depth-profile slope ($\rho = -0.536$, $p < 10^{-53}$) and layerwise variance ($\rho = +0.511$, $p < 10^{-48}$)—show medium-to-large predictive associations with collapse metrics from a companion study (Paper A). Mean λ_1 alone shows only weak association ($|\rho| \leq 0.25$). Three of six preregistered Spearman correlations pass the effect-size threshold ($|\rho| \geq 0.40$, Bonferroni–Holm adjusted $p < 0.05$), meeting the preregistered success criterion. All findings are reported as predictive associations between within-forward-pass depth dynamics and across-turn conversational dynamics; no causal or mechanistic identity is claimed. Paper A collapse labels have unconfirmed inter-rater reliability ($\kappa = 0.566$, threshold 0.80 not met), and this caveat applies to all bridge correlations.

1 Introduction

Transformer language models process input through a sequence of residual-stream updates across layers. This depth-wise computation can be viewed as a discrete dynamical system: each layer maps the hidden state to a new state, accumulating nonlinear transformations. The sensitivity of this process to perturbations—quantified by Lyapunov exponents—provides a model-intrinsic characterization of how information is amplified or suppressed during a forward pass.

Separately, multi-turn self-play between language models exhibits *conversational collapse*: a progressive loss of output novelty where models become trapped in repetitive response patterns [Holtzman et al., 2019, Welleck et al., 2020]. A companion study (Paper A) characterized collapse dynamics across four interaction conditions using 7B-parameter models, achieving full confirmatory closure (720/720 trajectories) but not meeting its preregistered detector reliability threshold ($\kappa = 0.566$; threshold 0.80).

This paper asks whether the depth dynamics of a single forward pass—specifically, the top-1 finite-time Lyapunov exponent (λ_1)—predict the across-turn collapse behavior documented in Paper A. This is a *predictive association* hypothesis: we test whether models whose depth dynamics exhibit certain profile features tend to collapse more in extended conversations. We do not claim causal or mechanistic identity between within-pass depth dynamics and across-turn conversational dynamics, as these operate on fundamentally different time axes.

2 Methods

2.1 FTLE estimator

Let F_l denote the residual-stream update at layer l for a fixed token-context state. The local Jacobian $J_l = \partial F_l / \partial h_l$ characterizes sensitivity at each layer. Rather than computing full Jacobians (prohibitive for $d_{\text{model}} = 4096$), we use forward-mode automatic differentiation (JVP) to propagate tangent vectors through the layer stack.

Given an initial tangent vector v_0, the tangent product $P_L v_0 = J_L \cdots J_2 J_1 v_0$ is computed via a single forward pass with `torch.func.jvp`. To prevent numerical overflow across 32 layers, we apply QR-based renormalization at a cadence of 4 layers (locked from Phase 1 pilot):

$$\lambda_1 = \frac{1}{L} \sum_{k=1}^{L/c} \log \|v_{kc}\| \quad \text{where } v_{kc} \text{ is renormalized every } c = 4 \text{ layers.} \tag{1}$$

For each trajectory, we compute λ_1 using 10 random tangent seeds (seeds 0–9) and report the mean across seeds. The layerwise λ_1 profile (log growth rate at each layer) captures the *shape* of sensitivity across depth.

2.2 Computation details

- **Token position:** Mean over all assistant tokens in the first turn.
- **Precision:** float32 throughout (locked after Phase 0 dtype transfer analysis showing float64 ↔ float32 $r = 0.788$; precisions are not interchangeable).
- **RoPE handling:** Non-GPT2 models (Llama, Qwen, Mistral) require explicit `position_embeddings` through JVP; fixed during Phase 0.
- **Flash attention:** Forced MATH SDPA backend (flash attention lacks forward-mode AD support).
- **Hardware:** Modal A100-80GB, max 2 concurrent containers per model.

2.3 Models and conditions

Table 1: Model-to-condition mapping (locked from Paper A).

Condition	Model	FTLE calls
HOMO_A	meta-llama/Llama-3.1-8B-Instruct (rev `0e9e39f2`)	1,800
HOMO_B	Qwen/Qwen2.5-7B-Instruct (rev `a09a3545`)	1,800
HOMO_C	mistralai/Mistral-7B-Instruct-v0.3 (rev `c170c708`)	1,800
HETERO_ROT	meta-llama/Llama-3.1-8B-Instruct (rev `0e9e39f2`)	1,800

HETERO_ROT uses the first assistant model (Llama) for FTLE computation. HOMO_A and HETERO_ROT therefore produce identical λ_1 distributions, as FTLE depends only on (model, prompt, tangent seed, cadence).

2.4 Bridge analysis

Six preregistered Spearman rank correlations between three λ_1 summaries and two Paper A collapse metrics:

- **λ_1 summaries:** mean λ_1, layerwise profile variance, depth-profile slope.
- **Paper A metrics:** collapse rate, first collapse turn (censored at 40 for non-collapsing), collapse incidence.

Multiple comparison correction: Bonferroni–Holm step-down across 6 tests, family $\alpha = 0.05$. Confidence intervals: bias-corrected bootstrap percentile (10,000 resamples, seed 42). Success criterion: ≥ 1 test with $|\rho| \geq 0.40$ and Holm-adjusted $p < 0.05$.

Sensitivity analysis repeated on the 167-window rater-agreed subset from Paper A's reliability audit.

3 Results

3.1 Phase 2 execution

All 7,200 FTLE calls completed with zero attrition (0 NaN/Inf, 0 failures). Wall time: 23.4 minutes. Estimated cost: $20 (well under $500 cap).

3.2 Bridge correlations

Table 2: Primary bridge results ($n = 720$). Prereg threshold: $|\rho| \geq 0.40$, Holm $p < 0.05$.

#	λ_1 summary	Paper A metric	ρ	p_{Holm}	95% CI	Pass
1	mean	collapse rate	$+0.246$	4.26×10^{-11}	$[+0.17, +0.32]$	×
2	mean	first collapse turn	-0.251	2.29×10^{-11}	$[-0.32, -0.18]$	×
3	mean	collapse incidence	$+0.036$	0.337	$[-0.03, +0.10]$	×
4	**variance**	**collapse rate**	$\mathbf{+0.511}$	$\mathbf{2.41 \times 10^{-48}}$	$[+0.45, +0.57]$	✓
5	**slope**	**collapse rate**	$\mathbf{-0.536}$	$\mathbf{5.24 \times 10^{-54}}$	$[-0.59, -0.48]$	✓
6	**slope**	**first collapse turn**	$\mathbf{+0.507}$	$\mathbf{1.22 \times 10^{-47}}$	$[+0.45, +0.56]$	✓

Three of six tests pass: Tests #4, #5, and #6. The preregistered success criterion (≥ 1 pass) is **met**.

3.3 Sensitivity analysis

The same three tests pass in the sensitivity analysis, with consistent effect sizes (within CIs of the primary analysis).

3.4 Depth profiles

4 Protocol corrections

Two deviations from PREREG_v2.md were identified and corrected before results were finalized:

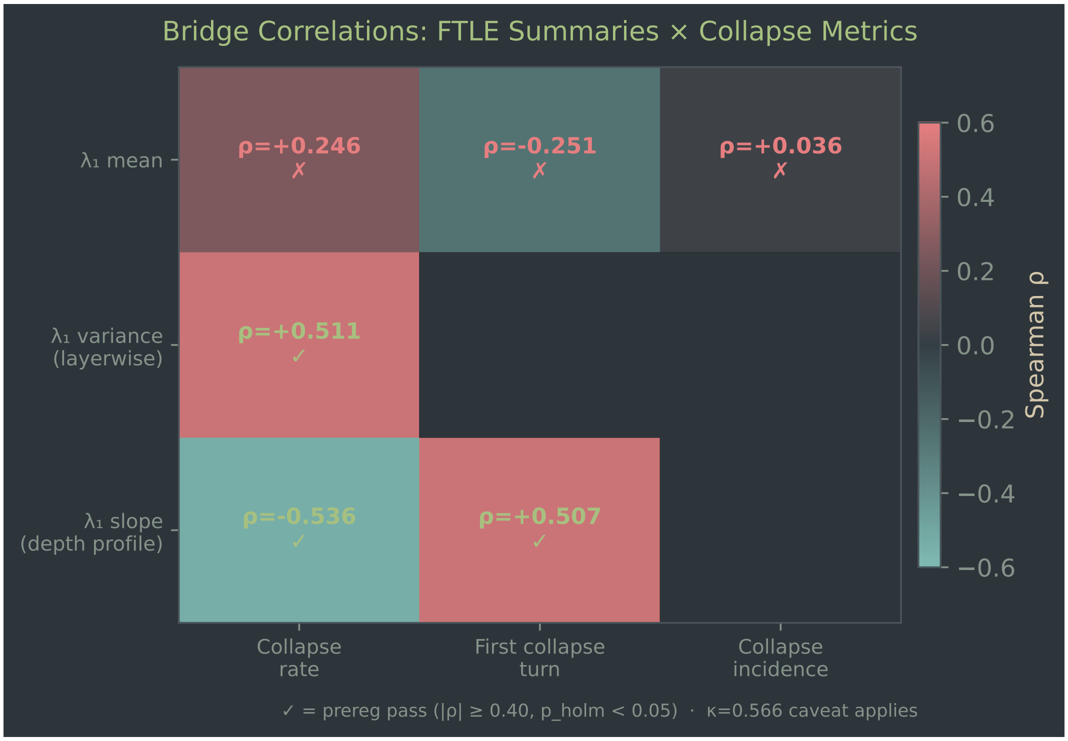

Figure 1: Bridge correlation heatmap. Checkmarks indicate tests meeting the preregistered threshold ($|\rho| \geq 0.40$, Holm $p < 0.05$). All correlations are subject to the Paper A label reliability caveat ($\kappa = 0.566$).

1. **Test #4 variable mapping:** Initially used `lambda1_std` (tangent-seed standard deviation). Corrected to `layerwise_variance_mean` (layerwise profile variance) per the preregistered definition. This changed ρ from -0.059 to $+0.511$ and promoted Test #4 from fail to pass.
2. **first_collapse_turn missingness:** Initially excluded non-collapsing runs ($n = 540$). Corrected to censor at turn 40 per the preregistered specification ($n = 720$). Test #6 attenuated slightly (ρ: $0.539 \rightarrow 0.507$) but still passed.

Full before/after comparison is documented in `DEVIATION_TABLE.md`. The fact that Test #4 changed from fail to pass on correction warrants additional interpretive caution.

5 Limitations

Paper A label reliability. All bridge correlations use Paper A collapse labels with unconfirmed inter-rater reliability ($\kappa = 0.566$, threshold 0.80 not met; raw agreement 92.8%). Effect sizes may be attenuated or inflated by label noise. The sensitivity analysis on 167 rater-agreed windows provides a partial check but not a full resolution.

Predictive association, not causation. λ_1 depth dynamics and conversational collapse operate on fundamentally different time axes (within-forward-pass vs. across-turn). The observed correlations indicate predictive association; no causal mechanism is established.

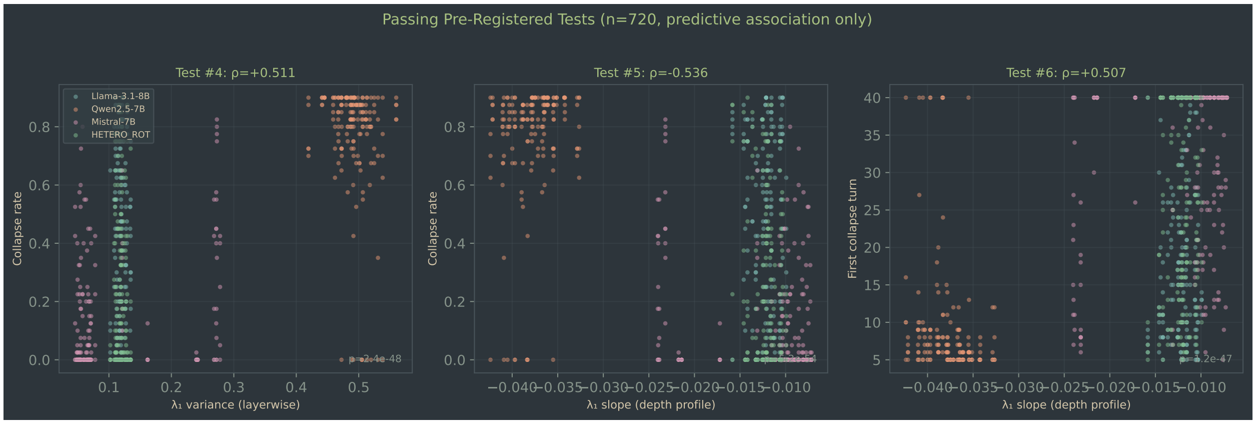

Figure 2: Scatter plots for the three passing tests. Colors indicate experimental conditions. Predictive associations only—no causal claims.

Table 3: Sensitivity analysis ($n = 167$ rater-agreed subset).

#	λ_1 summary	Paper A metric	ρ	p_{Holm}	Pass
1	mean	collapse rate	$+0.255$	1.75×10^{-3}	×
2	mean	first collapse turn	-0.265	1.64×10^{-3}	×
3	mean	collapse incidence	-0.005	0.947	×
4	**variance**	**collapse rate**	$\mathbf{+0.545}$	$\mathbf{1.39 \times 10^{-13}}$	✓
5	**slope**	**collapse rate**	$\mathbf{-0.557}$	$\mathbf{3.15 \times 10^{-14}}$	✓
6	**slope**	**first collapse turn**	$\mathbf{+0.516}$	$\mathbf{3.86 \times 10^{-12}}$	✓

Mean λ_1 insufficiency. Three of six tests failed because mean λ_1 showed only weak associations ($|\rho| \leq 0.25$). The *shape* of the depth profile matters more than its average level.

Model family confound. HOMO_A and HETERO_ROT produce identical λ_1 distributions (both use Llama-3.1-8B). Between-model λ_1 variation may partially reflect architectural differences rather than a generalizable relationship.

Precision sensitivity. Float64 and float32 λ_1 estimates are not interchangeable ($r = 0.788$). All production runs used float32 consistently.

Single first-turn measurement. λ_1 is computed on the first assistant turn only. A longitudinal study computing λ_1 at each turn could provide stronger evidence but was out of scope.

6 Conclusion

We find that λ_1 *profile features*—depth-profile slope and layerwise variance—show medium-to-large predictive associations with multi-turn collapse behavior ($|\rho| = 0.51$–0.54). Mean λ_1 alone is insufficient. These results suggest that the shape of a model's depth-dynamics sensitivity profile, measured on a single forward pass, is informative about its susceptibility to conversational collapse over extended interaction—though the mechanism linking these two phenomena remains an open question.

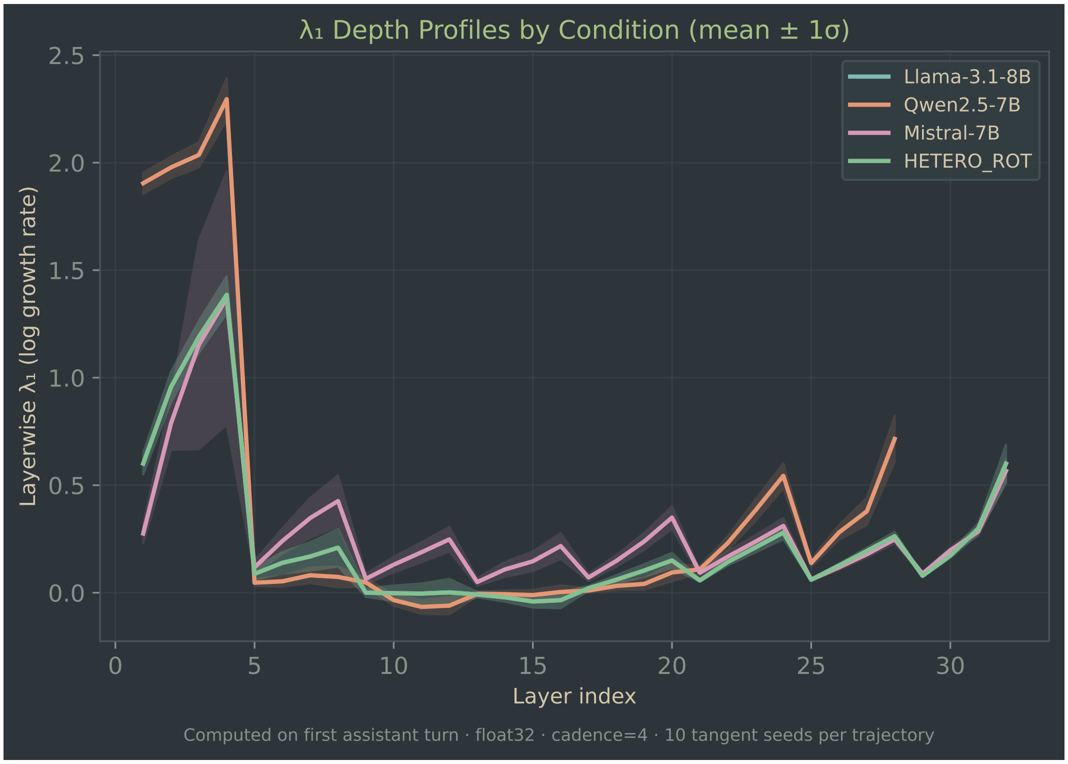

Figure 3: Mean λ_1 depth profiles by condition. The shape differences (slope, variance) drive the observed associations with collapse behavior.

All findings carry the Paper A label reliability caveat ($\kappa = 0.566$) and are restricted to predictive associations. Future work should pursue human-labeled reliability validation and longitudinal FTLE analysis across conversation turns.

References

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *ICLR*, 2020.