

The Multilingual Chaos Gap: Dynamical Warning Signals for Indic Safety Failures in Language Models

Vishwa Kumaresh

Draft: 2026-06-21

Abstract

Large language model safety behavior transfers unevenly across languages, especially for low-resource and code-mixed prompts. We ask whether this multilingual safety gap has a measurable dynamical signature inside the model during generation. Treating autoregressive decoding as a trajectory through residual-stream state space, we measure tokenizer fertility, recurrence structure, effective dimensionality, scalar local Lyapunov estimates, and early-warning-style trends in a safety-relevant readout-layer observable. In the current evidence package, two open 7B fallback models, Qwen2.5-7B-Instruct and Mistral-7B-Instruct-v0.3, were run on English plus Hindi, Tamil, and Bengali prompts with native, romanized, and code-mixed renderings. Indic renderings show large tokenizer-fertility increases and lower recurrence determinism relative to English in both models: conservative per-run bootstrap screens support all 9 fertility comparisons and 8 of 9 recurrence-determinism comparisons. A scalar-observable ablation over refusal, norm, and top-PC readouts keeps the strongest RQA-determinism support under the refusal-direction projection. A two-model readout-layer slice shows model-dependent depth behavior: Qwen localizes the recurrence signal to the configured mid layer and a later layer, while Mistral shows more persistent support across early, default, and late layers. Scalar Lyapunov estimates are mixed and do not establish a Lyapunov chaos-gap claim. Proxy safety diagnostics are also weak: static and fertility baselines beat the current EWS monitor under unvalidated proxy labels. We therefore report a constrained result: the current study validates a reproducible dynamical measurement pipeline and finds a robust trajectory-structure gap, while validated unsafe-token lead-time, monitor performance, and mitigation remain future work pending human labels and intended-model replication.

1 Introduction

Multilingual safety failures are an equity problem and a technical control problem. A user asking in Hindi, Tamil, Bengali, romanized text, or code-mixed South Asian usage should not receive weaker safeguards than a semantically equivalent English prompt. Prior work has shown large multilingual jailbreak and toxicity gaps [9, 10, 11, 12, 14], but most evaluations remain output-level: they tell us that a failure occurred, not whether the model entered a different internal generation regime before failing.

This paper studies a narrower mechanism question. Does the multilingual safety gap correspond to a change in the residual-stream trajectory during decoding? We model generation as a finite-time dynamical system, where each generated token yields a readout-layer hidden state. Safety-relevant behavior is then measured through trajectory statistics: recurrence, effective dimensionality, scalar instability, and early-warning-style trends.

The intended long-term contribution is a label-free internal monitor for multilingual safety failures. The current evidence is not yet a completed monitor result. Instead, it provides a repro-

ducible measurement pipeline and an evidence-limited finding: Indic and code-mixed renderings create systematically different input and trajectory structure from English in two open 7B models, while the current Lyapunov and proxy-monitor evidence is weak. This boundary is important. A paper-grade claim still requires intended Aya/Llama replication, validated human safety labels, token-level unsafe-onset labels, and final E2/E3 evaluation.

Our contributions in the current draft are:

- We implement a forward-pass-only pipeline for extracting residual-stream trajectories from quantized 7B instruction models and computing dynamical statistics with surrogate nulls and bootstrap screens.
- We report an E1 measurement over two open 7B fallback models and four languages, finding consistent Indic tokenizer-fertility increases and lower recurrence determinism relative to English.
- We add a scalar-observable ablation showing that the RQA-determinism result is strongest under the safety-grounded refusal-direction projection, weaker under top-PC, and weakest under norm.
- We separate negative and incomplete evidence from the main result: scalar Lyapunov estimates are mixed, proxy E3 monitor checks are weak, and validated safety/lead-time/mitigation results are not yet available.
- We release audit artifacts that map every major claim to evidence, including artifact, readiness, figure-slot, model-access, and goal-completion audits.

2 Related Work

Multilingual safety gaps. Low-resource and code-mixed prompts can bypass safety behavior more often than English prompts [9, 10, 11, 12, 13, 14]. This project focuses on Indic conditions because Hindi, Tamil, Bengali, romanized text, and code-mixing are practically important for South Asian users. The closest counter-hypothesis is that harmfulness may remain linearly represented while the model fails to act on it; we treat static separability as a baseline rather than assuming that representation disappears [7, 15].

Internal-state safety monitoring. Static hidden-state monitors and representation-engineering methods show that safety- or truth-related directions can be measured inside models [5, 6, 7, 8]. Concurrent trajectory monitors increasingly fire during decoding, but they usually use supervised probes or non-EWS geometric thresholds [18, 19, 17]. Our intended distinction is not “first temporal monitor.” It is the specific combination of label-free early-warning statistics, recurrence quantification, token-level lead-time, and multilingual mechanism testing.

Dynamical systems and early warning. Residual networks and transformers can be studied through dynamical-systems lenses [1, 2, 16]. Critical-slowness theory predicts rising lag-1 autocorrelation and variance near some tipping points [3, 4]; however, such signals are not universal. We therefore evaluate EWS statistics as hypotheses and include negative results rather than assuming that every safety failure is a tipping transition.

3 Method

3.1 Trajectory Extraction

For a prompt x , the model generates tokens $y_1, \dots, y_T$. At each step t , we record a single readout-layer residual-stream state $h_t \in \mathbb{R}^d$. Extraction uses one hooked layer, batch size 1, no gradient tracking, 4-bit model weights when needed, and CPU-side compressed trajectory storage. This design keeps the study within a 24 GB GPU budget.

3.2 Safety-Relevant Observable

The primary scalar observable is a projection onto a refusal direction:

$$z_t = r^\top h_t,$$

where r is estimated from harmful-versus-benign English trajectories. This observable is motivated by refusal-direction work [7]. The current evidence package includes scalar-observable ablations for norm and top-PC readouts and a two-model open-7B early/default/late readout-layer slice; intended-model layer robustness remains future work.

3.3 Dynamical Metrics

We compute five families of measurements. Tokenizer fertility controls for input fragmentation. Recurrence quantification measures whether the trajectory revisits similar states; the current headline recurrence statistic is determinism. Participation ratio estimates effective dimensionality. Scalar Rosenstein estimates provide a local instability diagnostic. Early-warning-style statistics use lag-1 autocorrelation and variance trends on the scalar observable. Phase-randomized surrogates and bootstrap confidence intervals are used as guardrails against overinterpreting noisy finite trajectories.

3.4 Validation And Auditing

Estimator validation is separated from empirical claims. The current validation artifact passes known-system checks for scalar estimators and direct twin-trajectory math, and a cached tiny-GPT-2 nnsight hook-injection smoke test confirms that perturbation plumbing can alter downstream hidden states. This does not constitute target-model twin-trajectory Lyapunov evidence.

Claim auditing is part of the method. The repository includes artifact, readiness, figure-slot, model-access, and goal-completion audits. These audits prevent proxy labels from being silently upgraded into validated attack-success, lead-time, monitor-performance, or mitigation claims.

4 Experiments

4.1 Setup

The current strongest runs use Qwen2.5-7B-Instruct and Mistral-7B-Instruct-v0.3. Each model is run over 50 harmful AdvBench seeds and 50 benign Alpaca seeds, rendered into English plus Hindi, Tamil, and Bengali native, romanized, and code-mixed conditions. This yields 1000 prompt rows per model and 2000 generations across the two-model evidence package. Non-English prompts use NLLB-200 translations. The intended Aya/Llama primary-model pair is not included because the current machine has invalid Hugging Face authentication and absent gated-model caches.

Table 1: Current artifact status for the two open 7B E1 runs. Counts are generated artifacts, not final safety labels.

Model	Prompt rows	Metric rows	Metric errors
Qwen2.5-7B-Instruct	1000	3000	0
Mistral-7B-Instruct-v0.3	1000	3000	0

Table 2: Current strongest E1 evidence. The table reports conservative sign-screen outcomes over 9 Indic rendering conditions.

Metric family	Direction vs. English	Supported conditions
Tokenizer fertility	Higher	9 / 9
RQA determinism	Lower	8 / 9
Scalar Rosenstein LE	Higher	0 / 9

4.2 E1: Trajectory-Structure Gap

The strongest current result is tokenizer fertility. Relative to English, fertility deltas are positive for all 9 Indic language/rendering conditions in both models, and all 9 pass the conservative all-run positive bootstrap screen. Large mean deltas appear for Tamil native (+9.833), Tamil romanized (+7.302), Bengali native (+5.579), and Hindi native (+4.121).

The strongest internal trajectory result is recurrence structure. RQA determinism decreases relative to English in both models for every Indic rendering. Mean deltas include Tamil native (-0.1185), Tamil code-mixed (-0.1076), Hindi romanized (-0.1044), and Bengali native (-0.0808). The conservative screen marks 8 of 9 recurrence-determinism comparisons as negative in all per-run bootstrap intervals.

Scalar Rosenstein Lyapunov estimates do not support the strongest chaos claim. Deltas are small-positive on average, but no Indic comparison has all-run positive-CI support and surrogate-null checks are mixed. We therefore describe LE as a fragile instability diagnostic, not as proof of a Lyapunov chaos gap.

4.3 Scalar-Observable Ablation

The observable ablation uses the existing matched-gap rows for refusal-direction projection, hidden-state norm, and trajectory-local top-PC projection. For harmful prompts, refusal-direction RQA determinism has all-run negative-CI support in 8 of 9 Indic conditions, top-PC RQA determinism in 5 of 9, and norm RQA determinism in 1 of 9. LE and AR(1) have no all-run CI support across the three scalar observables. This supports using the refusal-direction projection as the main readout in the current E1 measurement.

4.4 Readout-Layer Slice

A two-model depth slice compares six full-manifest E1 runs. For Qwen, harmful/refusal RQA determinism has 0 of 9 negative-CI Indic conditions at layer fraction 0.25, resolved to layer 7; 9 of 9 at the configured 0.55 layer, resolved to layer 15; and 9 of 9 at the 0.70 layer, resolved to layer 19. This suggests mid/late localization in Qwen. For Mistral, the same statistic has 7 of 9, 8 of 9, and 6 of 9 negative-CI Indic conditions at layer fractions 0.25, 0.55, and 0.70, resolved to layers 8, 17, and 22. This suggests a more depth-persistent but non-monotonic recurrence signal. Late-layer

Table 3: Goal-completion audit status. The external blocker is intended-model access.

Status	Count
PASS	9
PARTIAL	2
NOT_READY	4
BLOCKED_EXTERNAL	1

Rosenstein LE has 8 of 9 positive-CI conditions for Qwen but 0 of 9 for Mistral, so the layer slice strengthens the recurrence-structure result without settling the Lyapunov claim.

4.5 Proxy E2/E3 Diagnostics

Proxy safety diagnostics are intentionally separated from validated safety outcomes. Heuristic non-refusal proxy rates and provisional Qwen-rubric unsafe rates suggest several Indic conditions should be prioritized for review, but they are not attack-success rates. The benign-calibrated EWS monitor is weak under both proxy label sources: heuristic-refusal proxy AUC/TPR is 0.548/0.092 for Qwen and 0.480/0.077 for Mistral; Qwen-rubric proxy AUC/TPR is 0.467/0.072 for Qwen and 0.434/0.052 for Mistral.

Baseline comparisons further constrain the monitor story. Under pooled heuristic-refusal labels, the EWS score has raw proxy AUC 0.538, while post-hoc best-direction baselines are higher for static projection mean (0.935), static projection last (0.863), fertility (0.755), variance trend (0.717), and reversed RQA determinism (0.714). These proxy comparisons set the bar for future validated E3, but they do not establish deployable monitor performance.

4.6 Stochastic EWS Coverage

The stochastic sampled-rollout path now covers both 1000-row prompt sets. Qwen has mean AR(1) trend +0.0591 and mean variance trend -0.0910, with both-positive rate 0.257. Mistral has mean AR(1) trend -0.0833 and mean variance trend -0.0063, with both-positive rate 0.158. The mixed signs do not support an unqualified critical-slowness monitor claim.

4.7 Readiness And Missing Evidence

Table 3 summarizes the current completion state. The study has a reproducible E1 package and manuscript/proposal artifacts, but not a completed paper-grade E2/E3 study.

5 Discussion

The current results favor a modest conclusion. Indic and code-mixed prompts produce a robust input-side fragmentation gap and a consistent recurrence-structure shift in the readout-layer trajectory. This is compatible with a routing/control explanation of multilingual safety failures: static harmfulness information can remain available while the generation path is less likely to settle into a safe refusal mode. It is also compatible with a tokenizer-fragmentation explanation. The present evidence does not prove that safety failures are Lyapunov-chaotic tipping events.

The negative proxy-monitor results are useful. They prevent the paper from overclaiming a monitor before validated labels exist, and they identify the baselines future E3 must beat. In

particular, a label-free EWS monitor must outperform static projection and fertility controls under human-validated outcomes, not only under unvalidated proxy labels.

6 Limitations

The current study has four central limitations. First, the completed experiments use open Qwen/Mistral fallback models rather than the intended Aya/Llama pair. Second, semantic safety labels, translation-quality labels, and token-level unsafe-onset labels are blank. Third, target-model twin-trajectory Lyapunov experiments have not been run. Fourth, the current manuscript is not venue-ready: it requires human validation, intended-model replication, validated E2/E3, and final figure selection before top-tier submission.

7 Ethics And Safety

This work is defensive measurement research. It uses published red-team prompts and does not introduce new attack methods. The intended benefit is improved safety evaluation for low-resource and code-mixed users. Any release should preserve the audit artifacts and avoid reporting proxy labels as validated unsafe outcomes.

8 Conclusion

We introduced a reproducible dynamical measurement pipeline for multilingual LLM safety and ran it on two open 7B models across English, Hindi, Tamil, and Bengali prompt renderings. The current evidence supports a trajectory-structure gap, especially via tokenizer fertility and lower RQA determinism. It does not yet prove a Lyapunov chaos mechanism or a validated early-warning mitigation. The path to a top-tier paper is clear but incomplete: validate labels, run intended models, replace proxy E2/E3 with human-validated evaluations, and compare against strong static and trajectory-monitor baselines.

References

- [1] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural Ordinary Differential Equations. NeurIPS, 2018.
- [2] Boris Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on Transformers. arXiv:2312.10794, 2023/2024.
- [3] Marten Scheffer, Jordi Bascompte, William A. Brock, Victor Brovkin, Stephen R. Carpenter, Vasilis Dakos, Hermann Held, Egbert H. van Nes, Max Rietkerk, and George Sugihara. Early-warning signals for critical transitions. *Nature*, 461:53–59, 2009.
- [4] Thomas M. Bury, RI Sujith, Induja Pavithran, Marten Scheffer, Timothy M. Lenton, Madhur Anand, and Chris T. Bauch. Deep learning for early warning signals of tipping points. *PNAS*, 118, 2021.
- [5] Amos Azaria and Tom Mitchell. The Internal State of an LLM Knows When It’s Lying. Findings of EMNLP, 2023.

- [6] Samuel Marks and Max Tegmark. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. COLM, 2024.
- [7] Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in Language Models Is Mediated by a Single Direction. NeurIPS, 2024.
- [8] SafeSwitch. Hidden-state prober and refusal-head safety switching. arXiv:2502.01042, 2025.
- [9] Zheng-Xin Yong, Christina Menghini, and Stephen H. Bach. Low-Resource Languages Jailbreak GPT-4. ICLR, 2024.
- [10] Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. Multilingual Jailbreak Challenges in Large Language Models. ICLR, 2024.
- [11] Kang Min Yoo, Jinyoung Yang, and Jaesung Lee. Code-Switching Red-Teaming. ACL, 2025.
- [12] Harsh Aswal and Akshat Jaiswal. Haet Bhasha aur Diskrimineshun: Phonetic Perturbations in Hinglish. arXiv:2505.14226, 2025.
- [13] PolygloToxicityPrompts. Toxicity prompts for multilingual safety evaluation. COLM, 2024.
- [14] AI4Bharat. IndicLLMSuite: Sangraha, IndicAlign-Instruct, and IndicAlign-Toxic. ACL, 2024.
- [15] PolyRefuse. Refusal Direction is Universal Across Safety-Aligned Languages. arXiv:2505.17306, 2025.
- [16] Transformer Dynamics. A neuroscientific approach to interpretability of LLMs. arXiv:2502.12131, 2025.
- [17] Lyapunov Probes for Hallucination Detection. arXiv:2603.06081, 2026.
- [18] TrajGuard. Streaming hidden-state trajectory defense for jailbreak detection. arXiv:2604.07727, 2026.
- [19] Stop Early, Spend Less: Hidden-State Probes for Streaming Moderation. arXiv:2606.10487, 2026.