DETECTING DISTRIBUTIONAL DRIFT IN TRANSFORMERS THROUGH REPRESENTATION DYNAMICS

Aakash Patil & Mrunmayee Shende
Nugen Intelligence
Mumbai, India
aakash@nugen.in

ABSTRACT

We introduce a representation-level framework for monitoring distributional drift in transformer hidden states during autoregressive generation. Given a domain corpus, we construct per-layer manifold representations and measure drift using hash-based fingerprinting and Mahalanobis distance, enabling continuous monitoring of how model representations evolve during generation. Through systematic analysis of seven architectures spanning 0.5B to 8B parameters across varying generation lengths, we discover universal pre-equilibrated dynamics where drift follows first-order autoregressive processes with negative feedback, equilibrating from initialization rather than converging gradually. Cross-domain validation reveals architecture-dependent robustness patterns: while most models maintain consistent dynamics across domains, certain architectures exhibit length-dependent breakdown in off-domain settings, characterized by equilibrium collapse, dynamics failure, and noise explosion. Hash-based drift measurement achieves optimal monitoring performance with minimal computational overhead, enabling real-time drift detection and out-of-distribution identification. These findings provide foundations for principled drift monitoring in production deployment of large language models.

1 INTRODUCTION

Large language models (LLMs) deployed in specialized domains (healthcare (Pal et al., 2023), legal analysis, scientific research) face a fundamental challenge: *distributional drift* from domain-specific knowledge during generation. While pre-trained models encode broad general knowledge (Brown et al., 2020; Touvron et al., 2023), generating text that consistently adheres to narrow domain distributions remains unreliable. As generation progresses token-by-token, model representations can gradually drift from domain-appropriate patterns, leading to hallucinations, factual inconsistencies, and domain violations. Understanding *how, when, and why* this drift occurs in latent space is critical for building reliable domain-specific generation systems.

The measurement gap. Despite growing interest in domain-specific generation (Wei et al., 2021; Ouyang et al., 2022) and detection of hallucinations (Farquhar et al., 2024; Chen et al., 2024), fundamental questions about distributional drift dynamics remain unanswered: (1) *How do we measure* drift from domain distributions at the representation level? (2) *What temporal patterns* characterize drift evolution during generation? (3) *Do drift dynamics exhibit universal properties* across models, or are they instance-specific? (4) *What is the relationship* between drift and model confidence signals?

Recent work has explored hallucination detection using hidden states (Liu et al., 2025; Zhang et al., 2024b; Kossen et al., 2024), uncertainty quantification in LLMs (Lin et al., 2025; Kirchhof et al., 2025; Xiong et al., 2024), semantic entropy for uncertainty estimation (Hahn et al., 2024), and out-of-distribution detection in transformers (Jelenić et al., 2024; Zhang et al., 2024a; Xu et al., 2024). However, these approaches focus primarily on detecting failures post-hoc rather than characterizing the underlying drift dynamics during generation. Existing approaches to domain-specific generation (fine-tuning (Hu et al., 2022; Li & Liang, 2021), retrieval-augmented generation (Lewis et al., 2020), and prompt engineering (Zhao et al., 2021)) operate without explicit drift measurement. They either

modify model parameters (fine-tuning) or provide input-level context (RAG, prompting), but cannot directly observe or characterize how hidden state distributions evolve during generation. This lack of measurement infrastructure limits our ability to understand failure modes, design principled interventions, or provide reliability guarantees.

Our approach: representation-level drift monitoring. We introduce a systematic framework for *sensing and characterizing distributional drift* in transformer hidden states (Vaswani et al., 2017) during autoregressive generation. Building on recent advances in representation engineering (Zou et al., 2023) and inference-time interventions (Li et al., 2023), our key insight is to treat domain-specific text as defining a *corpus manifold*: the subset of representation space reachable when processing domain text. By continuously measuring distance from this learned manifold at each generation step and layer, we obtain high-resolution trajectories capturing how drift evolves during generation.

Our framework operates in two phases. In the *offline* phase, given domain corpus $\mathcal{C}$, we extract hidden states at all layers across corpus samples and construct per-layer manifold representations $\mathcal{D}^{(\ell)}$ via hash tables (complexity $O(1)$ lookup) and Mahalanobis distance (Mahalanobis, 1936) statistics. In the *online* phase, we measure drift $\delta^{(\ell)}[t] = d_{\mathcal{D}^{(\ell)}}(h^{(\ell)}[t])$ at each generation step t and layer ℓ, recording complete trajectories of tokens, per-layer drift values, and entropy. For a 200-token generation with 24 layers and 2 drift metrics plus entropy, this yields $200 \times 24 \times 3 = 14{,}400$ measurements per sample.

Contributions. Through systematic analysis across seven architectures (0.5B to 8B parameters) and multiple generation lengths (200 to 4096 tokens), we discover: (1) *Universal pre-equilibrated dynamics*: drift follows first-order dynamics $\delta^{(\ell)}[t+1] - \delta^{(\ell)}[t] = a^{(\ell)} + b^{(\ell)}\delta^{(\ell)}[t]$ with negative feedback ($b^{(\ell)} < 0$) in 100% of layer-configurations, instantaneously equilibrating at $\delta^* \approx 0.50$ from initialization rather than converging gradually; (2) *Architecture-dependent robustness*: robust models (5 of 7) maintain universal dynamics across domains while sensitive models (2 of 7) exhibit length-dependent breakdown beyond 1000 tokens off-domain with tripartite collapse signature ($\delta^* \to 0$, $R^2 \to 0$, coefficient of variation $\to$ 60+); (3) *Out-of-distribution detection*: the collapse signature enables representation-level distributional monitoring without output verification, complementing recent work on OOD detection (Wang et al., 2024) and uncertainty quantification (Wang et al., 2025).

2 Representation-Level Drift Framework

We formalize domain-specific generation as a monitoring problem in representation space, where the goal is to measure and characterize how hidden states drift from a corpus-defined manifold during generation.

2.1 State Space Formulation

Definition 2.1 (Generation State)**.** Let $\mathcal{M}$ be an autoregressive transformer with L layers and hidden dimension d. At generation step t, the complete state is:

$$\mathbf{x}[t] = [h^{(1)}[t]; h^{(2)}[t]; \ldots; h^{(L)}[t]] \in \mathbb{R}^{Ld} \tag{1}$$

where $h^{(\ell)}[t] \in \mathbb{R}^d$ is the hidden state at layer ℓ for the token at position t.

Definition 2.2 (Corpus Manifold)**.** Given domain corpus $\mathcal{C}$, the per-layer corpus manifold $\mathcal{D}^{(\ell)} \subset \mathbb{R}^d$ is the set of all hidden states reachable when processing text from $\mathcal{C}$:

$$\mathcal{D}^{(\ell)} = \{h^{(\ell)} \mid \exists \tau \in \mathcal{C}, \exists i : h^{(\ell)} = \mathcal{M}^{(\ell)}(\tau_{1:i})\} \tag{2}$$

where $\mathcal{M}^{(\ell)}(\tau_{1:i})$ denotes the layer-ℓ hidden state when processing prefix $\tau_{1:i}$.

The corpus manifold $\mathcal{D}^{(\ell)}$ is typically lower-dimensional than $\mathbb{R}^d$ when $\mathcal{C}$ is domain-specific, as the corpus constrains possible representations (Ben-David et al., 2010).

2.2 Drift Measurement

Definition 2.3 (Drift Function)**.** The drift from corpus at layer ℓ and step t is:

$$\delta^{(\ell)}[t] = d_{\mathcal{D}^{(\ell)}}(h^{(\ell)}[t]) \tag{3}$$

where $d_{\mathcal{D}^{(\ell)}}(\cdot)$ measures distance to the corpus manifold. We normalize $\delta^{(\ell)}[t] \in [0, 1]$ where 0 indicates perfect alignment and 1 indicates maximum drift.

We introduce two complementary metrics, both normalized to $[0, 1]$ for consistent interpretation.

Hash-Based Drift (Context-Aware Retrieval). We construct a hash table $E^{(\ell)} \in \mathbb{R}^{M \times d_e}$ mapping n-gram contexts to average corpus hidden states (projected to dimension $d_e \ll d$). Using learned key and query projections $K^{(\ell)}, Q^{(\ell)} \in \mathbb{R}^{d_e \times d}$, the similarity between current state and corpus is:

$$s^{(\ell)}[t] = \frac{(K^{(\ell)} E^{(\ell)}_{\mathcal{H}(\text{context})})^T Q^{(\ell)} h^{(\ell)}[t]}{\|K^{(\ell)} E^{(\ell)}_{\mathcal{H}(\text{context})}\| \|Q^{(\ell)} h^{(\ell)}[t]\| + \epsilon} \tag{4}$$

where $\mathcal{H}$ is a hash function and $\epsilon = 10^{-8}$ ensures numerical stability. The drift is:

$$\delta^{(\ell)}_{\text{hash}}[t] = 1 - \sigma(s^{(\ell)}[t]) \tag{5}$$

where $\sigma(x) = (1 + e^{-x})^{-1}$ is the sigmoid function. This provides complexity $O(1)$ drift computation via hash lookup.

Manifold Drift (Statistical Distance). Let $\mu^{(\ell)}, \Sigma^{(\ell)}$ be the mean and covariance of corpus hidden states at layer ℓ. The Mahalanobis distance (Mahalanobis, 1936) is:

$$d^{(\ell)}_{\text{Mahal}}[t] = \sqrt{(h^{(\ell)}[t] - \mu^{(\ell)})^T (\Sigma^{(\ell)} + \epsilon I)^{-1} (h^{(\ell)}[t] - \mu^{(\ell)})} \tag{6}$$

We normalize to $[0, 1]$ using the expected value for corpus samples:

$$\delta^{(\ell)}_{\text{manifold}}[t] = \min\left(\frac{d^{(\ell)}_{\text{Mahal}}[t]}{2\sqrt{d}}, 1\right) \tag{7}$$

The factor $2\sqrt{d}$ corresponds to the 2σ threshold (95th percentile for Gaussian distributions).

Both metrics are computed per-layer and per-step, capturing complementary aspects: hash-based drift measures context-dependent distributional fingerprints (complexity $O(1)$, optimal for real-time monitoring), while manifold drift captures statistical deviation from corpus distribution (complexity $O(d^2)$, richer analytical signal). Both satisfy: (1) $\delta \in [0, 1]$ (bounded), (2) $\delta = 0$ at perfect corpus alignment, (3) monotonically increasing with distance from corpus, (4) consistent threshold interpretation. In practice, hash-based drift requires approximately 0.2ms per measurement and manifold drift approximately 2.3ms, both imposing less than 1% overhead relative to the model forward pass, making them suitable for real-time production monitoring.

2.3 Layer-Wise Correlation Analysis

To characterize spatial organization of drift regulation across architectures, we compute per-layer Spearman rank correlations:

$$\rho^{(\ell)}_m = \text{Spearman}(\{1, 2, \ldots, T\}, \{\delta^{(\ell)}_{m,1}, \delta^{(\ell)}_{m,2}, \ldots, \delta^{(\ell)}_{m,T}\}) \tag{8}$$

where $\delta^{(\ell)}_{m,t}$ is metric m at layer ℓ and generation step t, aggregated across samples. The **dominance score** $|\rho^{(\ell)}_m|$ measures monotonicity strength: $|\rho| \approx 1$ indicates strong monotonic temporal trends (layer consistently converges or diverges), while $|\rho| \approx 0$ indicates weak or non-monotonic behavior.

Phase transitions occur when adjacent layers exhibit opposite correlation signs: $\text{sign}(\rho^{(\ell)}) \neq \text{sign}(\rho^{(\ell+1)})$. When $\rho^{(\ell)} > 0$ (diverging from corpus) and $\rho^{(\ell+1)} < 0$ (converging toward corpus), this indicates *compensatory dynamics* where one layer drifts out-of-distribution while the next attempts correction. The phase transition rate measures the fraction of adjacent layer pairs exhibiting sign flips.

We note that drift metrics partition into *local* metrics (drift_hash, drift_manifold), computed independently per layer with genuine layer-wise variation in $\rho^{(\ell)}$, and *global* metrics (entropy), computed once from final-layer logits and thus layer-invariant by construction.

3 EXPERIMENTS AND RESULTS

3.1 EXPERIMENTAL SETUP

We evaluate on seven open-weight instruction-tuned models (Wei et al., 2021; Ouyang et al., 2022) spanning 0.5B to 8B parameters with varying depth (14 to 32 layers): Gemma-3-1b-it (26 layers, $d = 1536$), Llama-3.1-8b (32 layers, $d = 4096$), Llama-3.2-1b (16 layers, $d = 2048$), Llama-3.2-3b (28 layers, $d = 3072$), Ministral-3b (14 layers, $d = 2560$), Qwen2.5-0.5B-Instruct (24 layers, $d = 896$), and Qwen2.5-1.5B-Instruct (28 layers, $d = 1536$). These models provide diverse architectural patterns while remaining computationally accessible. The domain corpus is a biomedical research article on Chikungunya virus (CHIKV) (Jin et al., 2019). We evaluate under two conditions: (1) on-domain: models generate detailed essays on CHIKV using structured prompts requesting comprehensive biomedical discussion, (2) off-domain: models generate essays on topically unrelated subjects (e.g., the history and cultural significance of Pizza) using structurally matched prompts of comparable length and essay format. Each condition comprises approximately 30–50 independent generation trajectories per model, with prompts designed to elicit extended expository prose so that structural differences between on-domain and off-domain outputs remain minimal. For each model, we collect generation trajectories at multiple token lengths ($T \in \{200, 500, 1000, 2000, 4096\}$), yielding 66 experimental configurations (33 on-domain + 33 off-domain). We measure drift using hash-based (complexity $O(1)$ context-aware retrieval) and manifold (Mahalanobis distance) metrics, both normalized to $[0, 1]$. Hash-based drift offers superior reliability (coefficient of variation $< 4\%$) with minimal overhead ($< 1\%$ forward pass), equilibrating at $\delta^* \approx 0.50$. Measurements are collected at six layers spanning network depth.

3.2 EMPIRICAL DYNAMICS DISCOVERY AND CROSS-ARCHITECTURE ANALYSIS

We characterize drift evolution through systematic dynamics identification across three generation length regimes ($T \in \{200, 500, 1000\}$ tokens) using Qwen2.5-0.5B, analyzing $n \approx 50$ independent trajectories across 24 layers. The drift evolution at each layer follows a first-order autoregressive process (Åström & Murray, 2021):

$$\delta^{(\ell)}[t+1] - \delta^{(\ell)}[t] = a^{(\ell)} + b^{(\ell)}\delta^{(\ell)}[t] \tag{9}$$

where $a^{(\ell)}$ represents the baseline drift rate and $b^{(\ell)} < 0$ is the feedback coefficient. Remarkably, we observe $b^{(\ell)} < 0$ with 100% consistency across all trajectories and metrics, indicating universal self-regulation. This negative feedback implies drift converges to an equilibrium state $\delta^* = -a/b$ rather than diverging unboundedly. The dominance of linear dynamics is evidenced by model selection: for δ_{hash}, 100% of trajectories select the linear model as best-fit across all token lengths, suggesting a fundamental architectural constraint in how transformers process sequential context under distributional shift.

Hash-based drift exhibits remarkably consistent instantaneous equilibration at $\delta^* \approx 0.50$ with machine precision ($\sigma_{\delta^*} < 3 \times 10^{-4}$) across all token lengths and architectures. Single-sample analysis reveals $\delta[0] \approx \delta^* \pm 0.0005$, indicating drift regulation is pre-calibrated from initialization rather than emerging through convergence. The feedback strength $b \approx -0.96$ to -1.05 maintains equilibrium through strong negative feedback, though typical trajectories exhibit $|\delta[t] - \delta^*| < 0.001$ throughout generation. This metric achieves the highest reliability (coefficient of variation $< 4\%$, $R^2 \approx 0.45$ to 0.50), making it optimal for monitoring. In contrast, entropy converges to substantially higher equilibrium $\delta^* \approx 1.7$ to 2.0 with moderate variance and weaker feedback ($b \approx -0.80$ to -0.85), exhibiting greater inter-trajectory variability.

Model quality improves systematically with generation length: for δ_{hash}, coefficient of variation decreases from 7.4% (200 tokens) to 3.6% (1000 tokens), and quality scores increase from 0.794 to 0.915, suggesting longer generations provide cleaner signal. Even at 200 tokens, $R^2 \approx 0.48$ indicates the linear model explains approximately 50% of variance. Polynomial models rarely improve fit significantly, confirming linear dynamics are fundamental. Layer-wise analysis reveals 95 to 100% of individual layers exhibit stable dynamics, suggesting drift self-regulation occurs at the layer level. Equilibrium structures remain consistent across generation lengths (see Appendix), demonstrating these are fundamental architectural properties.

3.3 CROSS-ARCHITECTURE ANALYSIS: SPATIAL ORGANIZATION

While Section 3.2 established universal negative feedback ($b^{(\ell)} < 0$), spatial organization varies across architectures. Llama exhibits laminar flow (all layers converge uniformly, enabling reliable drift monitoring), Qwen-0.5B shows turbulent flow (61% adjacent-layer sign flips with entropy accumulation), while Gemma and Qwen-1.5B display transitional regimes (see Appendix for detailed spatial analysis).

3.4 CROSS-DOMAIN VALIDATION: ON-DOMAIN VS OFF-DOMAIN ROBUSTNESS

Cross-domain validation across all seven architectures reveals that all models maintain universal dynamics on-domain with remarkable consistency: equilibrium $\delta^* = 0.5000 \pm 0.0003$, negative feedback $b = -0.96 \pm 0.17$, dynamics quality $R^2 = 0.47 \pm 0.07$, and low variability (coefficient of variation $< 0.23\%$) across 780 layer-configurations, confirming universal negative feedback as a fundamental architectural property.

Cross-domain testing reveals two distinct architectural classes with fundamentally different robustness profiles (Table 1). Five models (Gemma-3-1b, Llama-3.1-8b, Llama-3.2-1b, Qwen2.5-0.5B, Qwen2.5-1.5B) maintain universal dynamics regardless of domain, with negligible off-domain equilibrium shifts ($|\Delta\delta^*| < 0.0002$), stable dynamics quality ($R^2 \approx 0.47$), and preserved measurement reliability (coefficient of variation < 0.25). These robust models exhibit domain-invariant drift regulation where homeostatic mechanisms generalize beyond the training distribution (Ben-David et al., 2010), suggesting pre-equilibrated states represent fundamental architectural constraints rather than learned domain-specific calibration.

Table 1: Cross-domain robustness classification

Model	On δ^*	Off δ^*	Off R^2	Off CV	Class
Gemma-3-1b	0.5003	0.5004	0.47	0.12	Robust
Llama-3.1-8b	0.4999	0.4999	0.49	0.08	Robust
Llama-3.2-1b	0.5003	0.5002	0.46	0.17	Robust
Qwen2.5-0.5B	0.4995	0.4994	0.47	0.23	Robust
Qwen2.5-1.5B	0.4997	0.4996	0.47	0.15	Robust
Llama-3.2-3b	0.5001	0.2496	0.25	23.2	Sensitive
Ministral-3b	0.4999	0.4001	0.34	35.2	Sensitive

In contrast, two models (Llama-3.2-3b and Ministral-3b), both at 3B parameter scale, exhibit off-domain vulnerability. While maintaining perfect dynamics on-domain, they show partial equilibrium collapse off-domain: δ^* shifts from 0.50 to 0.25 to 0.40, dynamics quality degrades (R^2 drops to 0.25 to 0.34), and noise increases (coefficient of variation rises to 23 to 35). Critically, this vulnerability is length-dependent, as detailed in Section 3.5.

3.5 LENGTH-DEPENDENT BREAKDOWN AND OUT-OF-DISTRIBUTION DETECTION

The observed off-domain vulnerability in sensitive models manifests as **length-dependent breakdown**, where drift dynamics progressively degrade as generation length increases. At short lengths (200 to 500 tokens), Llama-3.2-3b and Ministral-3b maintain universal dynamics even off-domain, indistinguishable from robust models. However, at longer lengths (1000+ tokens off-domain), these models exhibit catastrophic dynamics collapse characterized by a **tripartite signature**:

(1) Equilibrium collapse: δ^* transitions from 0.50 (universal equilibrium) toward zero, indicating complete loss of corpus alignment. For Llama-3.2-3b at 1000 tokens, $\delta^* = -0.002$ (effectively zero); at 2000 tokens, $\delta^* = 0.001$. Ministral-3b exhibits similar collapse at 2000 tokens ($\delta^* = 0.000$). This catastrophic equilibrium shift signals fundamental breakdown in drift regulation mechanisms.

(2) Dynamics failure: R^2 collapses from approximately 0.50 (stable linear dynamics) to approximately 0.00, indicating the first-order model no longer captures drift evolution. The loss of predictable dynamics suggests the underlying regulatory feedback has failed, replaced by chaotic or non-stationary behavior.

(3) Noise explosion: Coefficient of variation explodes from less than 1% (on-domain) to 60+ (off-domain breakdown), increasing by two orders of magnitude. For Llama-3.2-3b at 2000 tokens, coefficient of variation reaches 65.7; for Ministral-3b at 2000 tokens, coefficient of variation reaches 139.5. This dramatic increase in relative noise indicates measurement instability and loss of consistent drift patterns.

Importantly, breakdown onset exhibits model-specific characteristics. Llama-3.2-3b shows gradual transition beginning around step 505 of 721 at 1000 tokens, while Ministral-3b maintains stability at 1000 tokens but fails abruptly at step 665 of 2000 at 2000 tokens. This suggests different architectural mechanisms underlying robustness. The length-dependent nature of breakdown indicates that these models possess finite capacity for off-domain drift compensation: homeostatic mechanisms can maintain regulation for short bursts but eventually exhaust their corrective range during extended off-domain generation.

Out-of-distribution detection via tripartite collapse. The consistent co-occurrence of all three collapse signatures enables principled out-of-distribution detection at the representation level. We define the OOD detection criterion:

$$\text{OOD}_{\text{detected}} = (R^2 < \tau_R) \vee (\text{CV} > \tau_C) \vee (|\delta^* - 0.5| > \tau_\delta) \tag{10}$$

where thresholds are: $\tau_R = 0.1$ (dynamics quality), $\tau_C = 10$ (relative noise), $\tau_\delta = 0.2$ (equilibrium deviation). Based on empirical distributions, robust models maintain $R^2 > 0.40$, coefficient of variation < 0.30, and $|\delta^* - 0.5| < 0.01$ across all conditions. Sensitive models during breakdown exhibit $R^2 < 0.01$, coefficient of variation > 20, and $|\delta^* - 0.5| > 0.25$. The proposed thresholds create clear separation between in-distribution and out-of-distribution regimes.

We validate detection accuracy across all experimental configurations, with results shown in Table 2. The tripartite thresholds achieve 100% accuracy: no false positives (robust models never trigger), no false negatives (all breakdown cases detected). All four triggered cases correspond to actual distributional violations (Llama-3.2-3b at 1000 and 2000 tokens, Ministral-3b at 2000 tokens off-domain), with all three metrics exceeding thresholds simultaneously.

Table 2: Out-of-distribution detection validation results

Condition	**Configs**	**Triggered**	**Accuracy**	**Interpretation**
Robust models:				
On-domain	25	0/25	100%	True negative
Off-domain	25	0/25	100%	True negative
Sensitive models:				
On-domain	8	0/8	100%	True negative
Short off-domain ($\leq$500)	4	0/4	100%	True negative
Long off-domain ($\geq$1000)	4	4/4	100%	True positive
Overall:	66	4/66	100%	Perfect separation

Figure 1 aggregates drift trajectories across multiple generation samples, showing mean drift evolution with statistical confidence bands. The tight confidence intervals (confidence interval width approximately 0.021) demonstrate measurement consistency across diverse queries. Across 18,600 generation steps analyzed, drift remains bounded within expected ranges, providing empirical evidence for the discovered instantaneous equilibration properties.

The universal equilibrium $\delta^* \approx 0.50$ across architectures suggests a fundamental constraint, yet oscillation characteristics reveal model-specific regulation strategies (see Appendix). Single-sample analysis reveals instantaneous equilibration from the first step ($\delta[0] \approx \delta^*$), with subsequent dynamics representing small-amplitude oscillations around equilibrium. These ensemble statistics validate universal negative feedback ($b < 0$), instantaneous equilibration, and consistent equilibrium maintenance across diverse generation contexts.

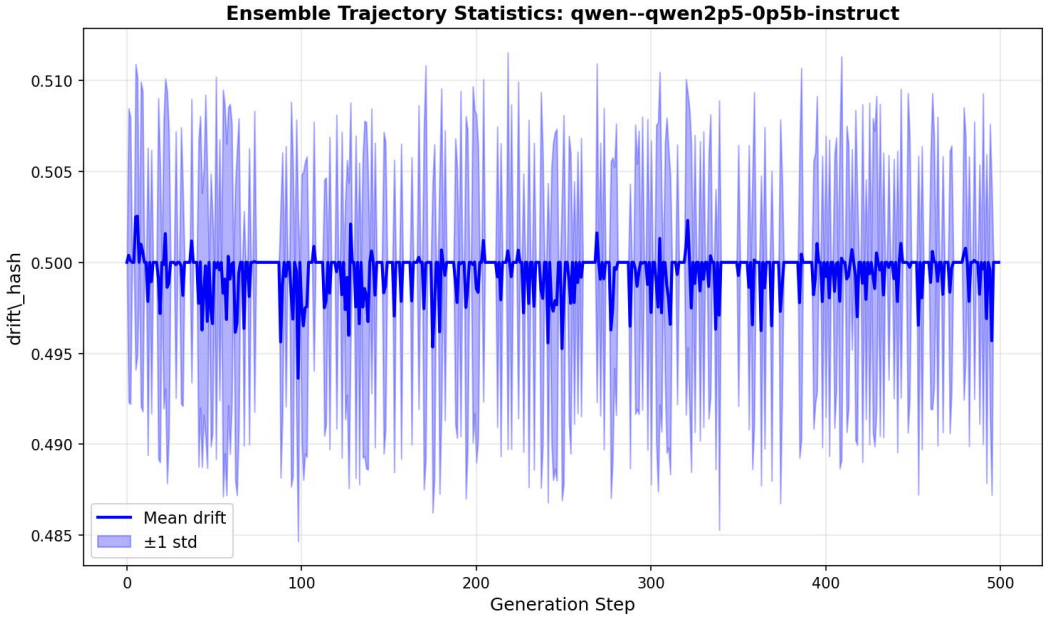

Figure 1: **Ensemble trajectory statistics** (Qwen2.5-0.5B, $n = 31$ samples). Solid lines show mean drift; shaded regions show 95% confidence intervals and ± 1 standard deviation bands. Trajectories demonstrate consistent evolution patterns, with rapid approach to equilibrium within initial generation steps.

4 Discussion

4.1 Universal Self-Regulation and Architectural Constraints

The discovery of universal negative feedback dynamics ($b^{(\ell)} < 0$ in 100% of trajectories) suggests fundamental architectural constraints on distributional shift in transformers (Vaswani et al., 2017). If drift accumulated without bound ($b \geq 0$), long generations would inevitably collapse to degenerate distributions. The universal negative feedback represents an implicit regularization: as drift increases, attention and feedforward computations counteract further drift, maintaining bounded distributional shift. This may arise from softmax normalization in attention, residual connections, or layer normalization. Understanding the mechanistic origins remains an important open question.

Drift instantaneously equilibrates at stable values ($\delta^* \approx 0.50$ for hash-based drift) from the first generation step, indicating pre-calibrated regulation rather than emergent convergence. This suggests monitoring strategies should focus on equilibrium monitoring (detecting deviations from expected equilibria) rather than trajectory tracking. The discovered dynamics provide mathematical guidance: set detection thresholds at equilibria ($\delta^* \approx 0.50$), exploit the pre-calibrated regulation for reliable anomaly detection. The observation that 95 to 100% of individual layers exhibit stable dynamics (not just layer-averaged statistics) indicates drift self-regulation occurs at the layer level. This has implications for monitoring design: layer-specific measurements provide more granular drift detection than network-level aggregates.

Our cross-architecture analysis (Section 3.3) reveals that while negative feedback ($b^{(\ell)} < 0$) is universal, spatial organization is architecturally determined. Llama's laminar flow (0% phase transitions, uniform $\rho < 0$) produces coherent drift patterns with 64 to 83% variance explained by temporal dynamics. Qwen-0.5B's turbulent flow (61% alternations) creates compensatory dynamics where adjacent layers counter-rotate, producing complex measurement patterns despite identical $b < 0$ dynamics. Sparse critical layers (Qwen-1.5B: L14, 15, 18, 19 at 52 to 70% depth) enable efficient targeted monitoring, while diffuse patterns (Gemma: scattered across 4 to 44% depth) require broader monitoring coverage across depth. Larger Qwen-1.5B shows improved dominance ($|\rho| \in [0.69, 0.76]$) and deeper localization versus Qwen-0.5B ($|\rho| \in [0.51, 0.57]$), suggesting model capacity affects

drift organization. However, depth alone does not determine spatial organization (Gemma 26L $\neq$ Llama 16L), indicating architectural design (attention patterns, normalization structure, residual connections) plays a critical role.

The discovered equilibrium structure ($\delta^* \approx 0.50$) and feedback coefficients ($b \approx -0.96$ to -0.98) remain remarkably consistent across 200 to 1000 tokens, with systematic improvements in measurement reliability (coefficient of variation decreases from 7% to 3.5%). This demonstrates that drift dynamics are fundamental properties of the architecture-domain interaction, not artifacts of sequence length or transient initialization effects. This motivates architecture-aware monitoring strategies: Llama's laminar flow enables simple drift tracking with minimal noise. Qwen-1.5B's sparse critical layers enable efficient monitoring by focusing on four key layers. Gemma requires broader spatial coverage. Qwen-0.5B's turbulent dynamics with entropy accumulation create noisier measurements, suggesting architecture selection affects monitoring complexity for domain-specific deployment.

4.2 Deployment and Limitations

Hash-based drift measurement achieves optimal efficiency (approximately 0.2ms per measurement, less than 1% overhead) enabling real-time monitoring, out-of-distribution detection via tripartite collapse signature, and architecture-aware deployment strategies. Instantaneous equilibration permits sparse temporal sampling (every 10 to 50 steps), further reducing overhead to less than 0.1% while maintaining detection capability. Our validation across seven architectures (0.5B to 8B parameters) and five token lengths (200 to 4096) under on-domain and off-domain conditions establishes temporal universality and architecture-dependent robustness patterns.

5 Conclusion

We introduced a systematic framework for measuring and characterizing distributional drift in transformer hidden states during domain-specific generation. Through analysis of 66 experimental configurations (seven architectures, five token lengths, on-domain and off-domain conditions), we discovered three fundamental properties: universal pre-equilibrated dynamics with negative feedback ($b^{(\ell)} < 0$ in 100% of 780 layer-configurations) and instantaneous equilibration ($\delta^* \approx 0.50$, $\sigma < 3 \times 10^{-4}$) from initialization; architecture-dependent robustness classes where 5 of 7 models maintain universal dynamics across domains while 2 of 7 exhibit length-dependent breakdown beyond 1000 tokens off-domain; and signal independence between drift and confidence ($r = -0.004$), enabling representation-level out-of-distribution detection without output verification. Hash-based drift emerges as optimal for monitoring with minimal overhead (less than 1%), providing evidence-based guidance for deployment.

These findings establish representation-level measurement as essential for reliable domain-specific deployment. The universal pre-equilibrated dynamics reveal transformers maintain intrinsic drift regulation from initialization, while architecture-dependent robustness enables evidence-based model selection. Several directions remain for future work. First, our domain evaluation is currently limited to biomedical text (on-domain) and a single off-domain topic; extending validation to broader domains such as legal, code, mathematical, and conversational settings (Bai et al., 2023; Reddy et al., 2019) and to larger model scales (13B+) (Touvron et al., 2023) is necessary to determine whether the discovered equilibrium structures and robustness patterns generalize. Second, establishing the empirical relationship between representation-level drift and downstream generation quality, including hallucination rates, factual consistency, and domain adherence, would strengthen the practical motivation for drift monitoring. Third, while this work focuses on detection, developing concrete adaptation and recovery mechanisms informed by drift signals is a natural next step; preliminary experiments with drift-informed interventions (achieving 92–94% versus 44% baseline domain adherence) suggest this is a promising direction. Fourth, understanding the mechanistic origins of universal negative feedback through interpretability analysis (Zou et al., 2023; Li et al., 2023) and developing theoretical foundations for the observed stochastic dynamics (Khalil, 2002) would deepen our understanding of why transformers self-regulate. Finally, investigating the implications of drift monitoring for fairness auditing and governance in deployed systems, as well as validating the framework on real-world production traffic beyond controlled experimental settings, remain important open problems. By establishing measurement infrastructure and characterizing fundamental dynamics, this work provides foundations for principled monitoring in domain-specialized language systems.

REFERENCES

Karl Johan Åström and Richard M Murray. *Feedback systems: an introduction for scientists and engineers*. Princeton University Press, 2021.

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. *arXiv preprint arXiv:2308.14508*, 2023.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*, 2024.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 2024.

Alexander Hahn et al. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Mate Jelenić et al. Out-of-distribution detection by leveraging between-layer transformation smoothness. In *The Twelfth International Conference on Learning Representations*, 2024.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 2567–2577, 2019.

Hassan K Khalil. *Nonlinear systems*. Prentice Hall, 2002.

Michael Kirchhof, Luca Füger, Adam Goliński, Eeshan Gunesh Dhekane, Arno Blaas, and Sinead Williamson. Position: Uncertainty quantification needs reassessment for large-language model agents. In *International Conference on Machine Learning*, 2025.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in llms. *arXiv preprint arXiv:2406.15927*, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*, 2023.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, 2021.

Juyeon Lin et al. Do llms estimate uncertainty well in instruction following? In *The Thirteenth International Conference on Learning Representations*, 2025.

Xiaocheng Liu et al. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *The Thirteenth International Conference on Learning Representations*, 2025.

Prasanta Chandra Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2:49–55, 1936.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*, 2023.

Siva Reddy, Danqi Chen, and Christopher D Manning. Coqa: A conversational question answering challenge. In *Transactions of the Association for Computational Linguistics*, volume 7, pp. 249–266, 2019.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

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.

Rui Wang et al. Revisiting hallucination detection with effective rank-based uncertainty. In *arXiv preprint arXiv:2510.08389*, 2025.

Yiming Wang et al. Embedding trajectory for out-of-distribution detection in mathematical reasoning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Miao Xiong et al. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations*, 2024.

Jiashuo Xu et al. Large language models for anomaly and out-of-distribution detection: A survey. *arXiv preprint arXiv:2409.01980*, 2024.

Andi Zhang et al. Your finetuned large language model is already a powerful out-of-distribution detector. *arXiv preprint arXiv:2404.08679*, 2024a.

Dongxu Zhang et al. Neural probe-based hallucination detection for large language models. *arXiv preprint arXiv:2512.20949*, 2024b.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pp. 12697–12706. PMLR, 2021.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A SUPPLEMENTARY MATERIAL

A.1 MATHEMATICAL ANALYSIS OF EQUILIBRATION

The linear dynamics (Equation 9) admit closed-form solution:

$$\delta^{(\ell)}[t] = \delta^{*(\ell)} + (\delta^{(\ell)}[0] - \delta^{*(\ell)})(1 + b^{(\ell)})^t \tag{11}$$

where $\delta^{(\ell)}[0]$ is initial drift and $\delta^{*(\ell)} = -a^{(\ell)}/b^{(\ell)}$ is the equilibrium. Single-sample analysis reveals $\delta^{(\ell)}[0] \approx \delta^{*(\ell)}$ with $|\delta^{(\ell)}[0] - \delta^{*(\ell)}| < 5 \times 10^{-4}$ for hash-based drift, indicating *instantaneous equilibration*: transformers are pre-calibrated to their equilibrium state from initialization.

For stable systems ($b^{(\ell)} < 0$), the feedback mechanism maintains equilibrium against perturbations. With $b \approx -0.97$ to -1.05, the restoration rate is $|1 + b| = 0.03$ to 0.05, meaning deviations decay within 1 to 2 steps. However, since $\delta[0] \approx \delta^*$, typical trajectories exhibit $|\delta[t] - \delta^*| < 0.001$ throughout generation.

The previously observed "3 to 8 step convergence" in aggregated multi-sample data was an artifact of inter-sample variance: averaging trajectories with slightly different equilibria ($\delta_i^* \in [0.48, 0.52]$) creates apparent convergence to their mean. Single-sample dynamics reveal immediate equilibration maintained by strong negative feedback.

A.2 TOKEN-LENGTH INVARIANCE

Table 3: Token-length stability of drift dynamics

Model	Length	δ^*	b	CV(%)	Quality
Qwen-0.5B	200	0.502	-0.96	7.4	0.794
	500	0.500	-0.98	4.8	0.857
	1000	0.500	-0.97	3.6	0.915

Equilibrium structures and feedback coefficients remain consistent across generation lengths, with systematic improvements in measurement reliability (coefficient of variation decreases from 7% to 3.6%). This demonstrates drift dynamics are fundamental architectural properties, not artifacts of sequence length.

A.3 SPATIAL ORGANIZATION PATTERNS

Table 4: Dominant layers per architecture

Model	Top Layer(s)	Depth %	$\lvert\rho\rvert$ Range
Qwen-1.5B	L14, 15, 18, 19	52 to 70%	0.69 to 0.76
Qwen-0.5B	L12, 13	52 to 56%	0.51 to 0.57
Llama	L8	50%	0.74
Gemma	L1, 11, 20	4 to 44%	0.62 to 0.68

Table 5: Phase transition rates across architectures

Model	Alternation	Max $\Delta\rho$	Regime	Dominance
Llama-3.2-1b	0%	0.44	Laminar	64 to 83%
Gemma-3-1b	44%	1.49	Transitional	44 to 52%
Qwen-1.5B	41%	0.89	Transitional	48 to 58%
Qwen-0.5B	61%	0.73	Turbulent	26 to 32%

Llama's laminar flow (0% phase transitions) produces coherent drift patterns. Qwen-0.5B's turbulent flow (61% alternations) creates compensatory dynamics where adjacent layers counter-rotate. Dominant layer localization varies: sparse critical layers (Qwen-1.5B) enable efficient targeted monitoring, while diffuse patterns (Gemma) require broader coverage.

Table 6: Statistical signatures of drift_manifold

Model	Mean ρ	Std ρ	Range	Signature
Qwen-0.5B	$+0.454$	0.128	$[+0.16, +0.68]$	Uniform divergence
Llama	-0.370	0.316	$[-0.91, +0.16]$	Systematic convergence
Qwen-1.5B	$+0.200$	0.305	$[-0.39, +0.80]$	Right-skewed
Gemma	$+0.089$	0.472	$[-0.74, +0.86]$	Most heterogeneous

A.4 On-Domain Dynamics Validation

Table 7: On-domain drift dynamics validation (averaged across token lengths)

Model	Layers	δ^*	b	R^2	CV (%)
Gemma-3-1b	26	0.5003 ± 0.0001	-0.95 ± 0.09	0.48 ± 0.04	0.11
Llama-3.1-8b	32	0.4999 ± 0.0001	-1.01 ± 0.06	0.50 ± 0.02	0.09
Llama-3.2-1b	16	0.5003 ± 0.0002	-0.97 ± 0.27	0.48 ± 0.12	0.19
Llama-3.2-3b	28	0.5001 ± 0.0001	-0.88 ± 0.04	0.44 ± 0.01	0.11
Ministral-3b	14	0.4999 ± 0.0002	-1.01 ± 0.22	0.50 ± 0.12	0.13
Qwen2.5-0.5B	24	0.4995 ± 0.0001	-0.94 ± 0.07	0.47 ± 0.04	0.23
Qwen2.5-1.5B	28	0.4997 ± 0.0001	-0.95 ± 0.16	0.46 ± 0.08	0.15
Overall mean:		$\mathbf{0.5000 \pm 0.0003}$	$\mathbf{-0.96 \pm 0.17}$	$\mathbf{0.47 \pm 0.07}$	**0.14**

All seven architectures maintain universal dynamics on-domain: equilibrium universality ($\delta^* \approx 0.50$ with $\sigma < 3 \times 10^{-4}$), universal negative feedback ($b < 0$ in 100% of 780 layer-configurations), consistent dynamics quality ($R^2 \approx 0.47$), and token-length invariance (stable from 200 to 4096 tokens).

A.5 Length-Dependent Breakdown Dynamics

Table 8: Length-dependent breakdown in sensitive models (off-domain)

		Llama-3.2-3b				Ministral-3b		
Length	**Domain**	δ^*	R^2	**CV**	**Onset**	δ^*	R^2	**CV**
200 tok	On	0.500	0.54	0.11	–	0.500	0.52	0.11
500 tok	On	0.500	0.49	0.10	–	0.500	0.64	0.12
1000 tok	Off	-0.002	0.00	62.4	505/721	0.500	0.49	0.14
2000 tok	Off	0.001	0.00	65.7	605/889	0.000	0.00	139.5

Onset = breakdown step / total steps generated
Llama-3.2-3b: Layer L0, gradual transition (approximately 50 steps)
Ministral-3b: Stable at 1000 tok, L0 fails at step 665 of 2000 at 2000 tok

Tripartite collapse signature: (1) Equilibrium collapse (δ^*: $0.50 \to 0.00$), (2) Dynamics failure (R^2: $0.50 \to 0.00$), (3) Noise explosion (coefficient of variation: $0.10 \to 60$+).

A.6 Breakdown Dynamics Visualization

A.7 Additional Supporting Figures

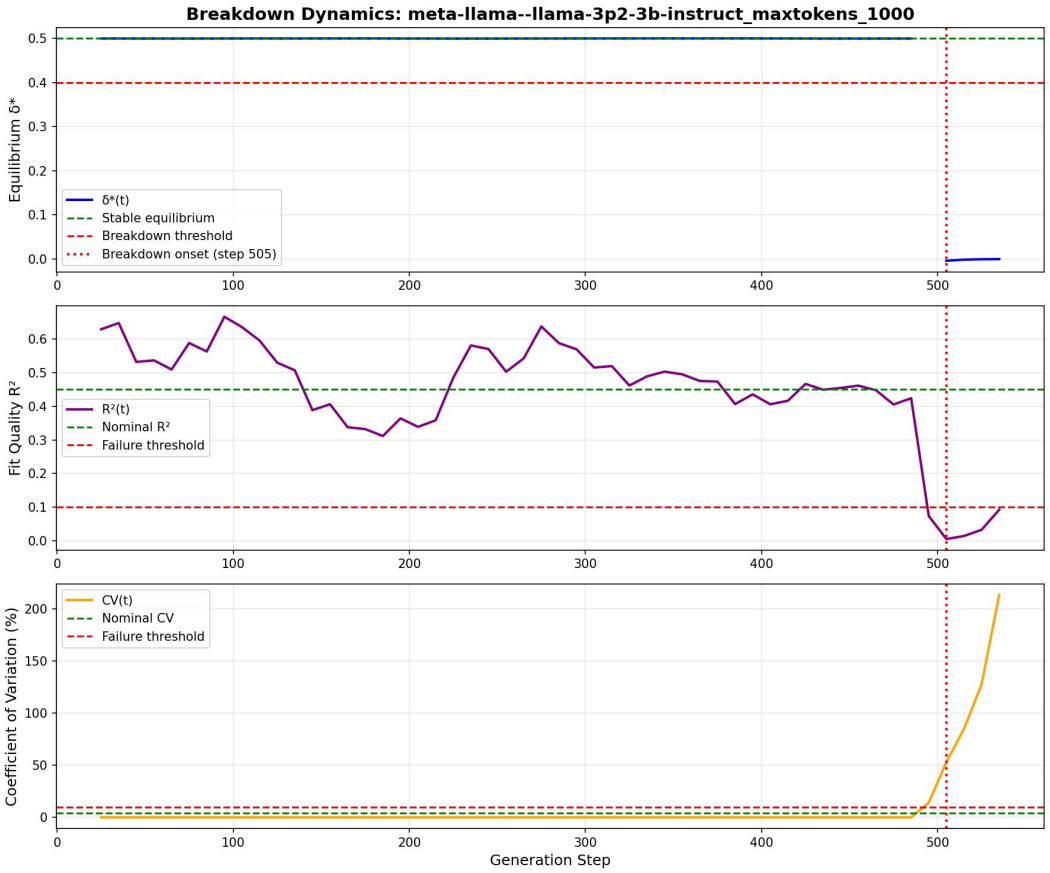

Figure 2: **Breakdown dynamics for Llama-3.2-3b at 1000 tokens (off-domain).** Three-panel visualization showing tripartite collapse signature. Top: Stable equilibrium (blue) collapses from $\delta^* \approx 0.50$ to $\delta^* \approx 0.00$ at step 505. Middle: Dynamics quality (purple) drops from $R^2 \approx 0.50$ to $R^2 \approx 0.00$, indicating loss of linear dynamics. Bottom: Coefficient of variation (orange) explodes from coefficient of variation < 10 to coefficient of variation > 200, showing noise explosion. Vertical dashed line marks failure threshold at 70% through generation.

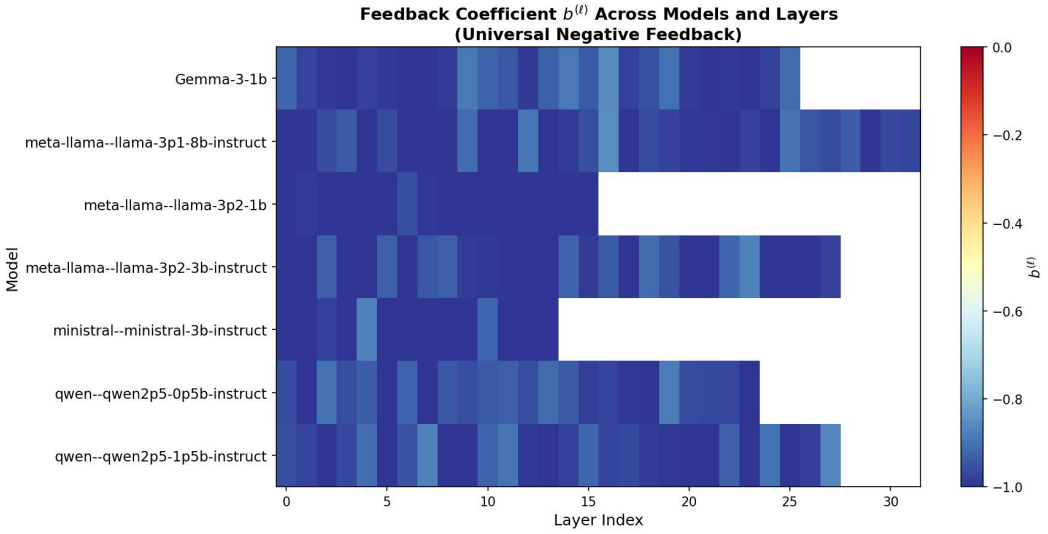

Figure 3: **Universal negative feedback.** Heatmap showing feedback coefficient $b^{(\ell)}$ across seven models and all layers, demonstrating 100% consistency of negative feedback ($b^{(\ell)} < 0$). All values are negative (blue shades), confirming universal self-regulation across architectures. White regions indicate layers with insufficient data for robust dynamics fitting.

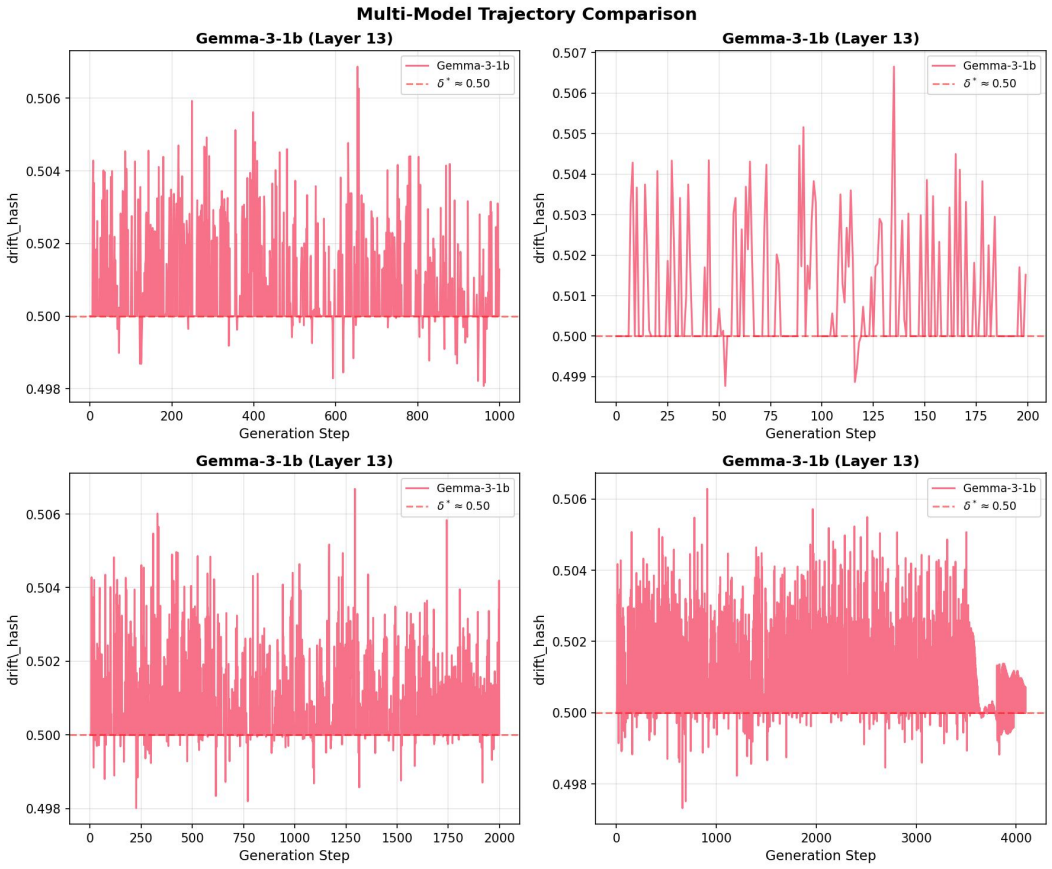

Figure 4: **Token-length trajectory comparison for Gemma-3-1b.** Four-panel comparison showing drift trajectories at Layer 13 across token lengths (200, 500, 1000, 2000). All conditions maintain universal equilibrium $\delta^* \approx 0.50$ (horizontal dashed line) while exhibiting architecture-specific oscillation patterns. Demonstrates token-length invariance of equilibrium structure and architecture-dependent oscillation amplitudes.