Numerical Instability and Chaos: Quantifying the Unpredictability of Large Language Models

Chashi Mahiul Islam[1], Alan Villarreal[1], Mao Nishino[2], Shaeke Salman[1], Xiuwen Liu[1]
[1]*Department of Computer Science, Florida State University, Tallahassee, USA*
[2]*Department of Mathematics, Florida State University, Tallahassee, USA*
ci20l@fsu.edu, adv23a@fsu.edu, mnishino@fsu.edu, salman@cs.fsu.edu, xliu@fsu.edu

***Abstract*—As Large Language Models (LLMs) are increasingly integrated into agentic workflows, their unpredictability stemming from numerical instability has emerged as a critical reliability issue. While recent studies have demonstrated the significant downstream effects of these instabilities, the root causes and underlying mechanisms remain poorly understood. In this paper, we present a rigorous analysis of how unpredictability is rooted in the finite numerical precision of floating-point representations, tracking how rounding errors propagate, amplify, or dissipate through Transformer computation layers. Specifically, we identify a chaotic "avalanche effect" in the early layers, where minor perturbations trigger binary outcomes: either rapid amplification or complete attenuation. Beyond specific error instances, we demonstrate that LLMs exhibit universal, scale-dependent chaotic behaviors characterized by three distinct regimes: 1) a stable regime, where perturbations fall below an input-dependent threshold and vanish, resulting in constant outputs; 2) a chaotic regime, where rounding errors dominate and drive output divergence; and 3) a signal-dominated regime, where true input variations override numerical noise. We validate these findings extensively across multiple datasets and model architectures.**

***Index Terms*—Large Language Models, Numerical Instability, Chaotic Dynamics, Floating-Point Precision, Rounding Errors, Robustness**

I. Introduction

The deployment of Large Language Models (LLMs) has rapidly evolved from isolated inference systems to complex, distributed multi-agent architectures where multiple AI agents collaborate to solve sophisticated tasks [1], [2]. Multi-agent LLM systems improve performance through collaborative problem-solving [3], [4], but they exhibit high failure rates that cannot be attributed solely to algorithmic limitations. Wu et al. [4] report that AutoGen-based workflows fail to converge on 23% of collaborative tasks, with agents producing contradictory outputs despite identical prompts. Similarly, Hong et al. [5] document non-reproducible outputs in 31% of MetaGPT planning tasks across identical hardware configurations. These failures share a common signature: unpredictability that persists even with fixed random seeds, manifesting most severely when agents exchange intermediate representations.

We hypothesize that a significant fraction of these multi-agent failures stem from numerical instability induced by floating-point computation across heterogeneous infrastructure. In modern LLM deployments, floating-point arithmetic is neither associative nor deterministic across diverse hardware [6]. When agents communicate by exchanging embeddings across cloud environments with diverse GPU architectures, numerical discrepancies arise from non-deterministic parallel reduction orders [7], hardware-specific implementations [8], and accumulation of rounding errors through deep network layers. The Oak Ridge National Laboratory documents that "the variability from non-deterministic reductions will produce unique and non-reproducible results with each run using identical inputs" [9]. Zhuang et al. [10] demonstrate that GPU-based training produces non-reproducible results in 100% of trials across different hardware.

This impact is amplified by the extreme sensitivity of LLMs. Hahn [11] proves that self-attention mechanisms have condition numbers that grow exponentially with sequence length. Kim et al. [12] demonstrate empirically that LLMs are "extremely sensitive to assigned prompts," with 13.78% of correct answers becoming incorrect solely due to minor prompt perturbations. This sensitivity extends to embeddings: Vijayaraj [13] shows that perturbations of magnitude 10^{-3} in normalized embedding space cause diagnostic prediction flips in 12.4% of clinical cases.

Prior work has documented numerical instability in deep learning, and multi-agent system fragility [4], [9], [10], [14], but a gap remains: *we lack a principled understanding of how floating-point rounding errors interact with LLM computation to produce unpredictable outputs*. Existing work treats numerical instability as an engineering nuisance to be mitigated through deterministic execution modes [15] or higher precision arithmetic, without characterizing the underlying dynamics.

This paper addresses this gap through rigorous analysis of numerical stability in LLM inference. Our contributions are:

1) **Identification of Chaotic Dynamics.** We demonstrate that LLMs exhibit chaotic behavior where perturbations at the scale of floating-point epsilon ($\sim 10^{-14}$) either amplify exponentially or attenuate completely within early Transformer layers, with directional absolute condition numbers exceeding 10^6.
2) **Characterization of Three Stability Regimes.** We identify three distinct operating regimes: *Constant Regions* where output is bitwise-constant; *Chaotic Regions* with extreme sensitivity; and *Signal-Dominated Regions* where input variations override numerical noise.
3) **Empirical Validation.** We validate findings across multiple combinations of settings, including architectures (Llama-3.1-8B, GPT-OSS-20B), datasets (TruthfulQA,

AdvBench), and floating-point precision (BFloat16, FP32, and FP64), demonstrating universal phenomena rather than model-specific artifacts.

As LLMs are integrated into safety-critical applications [4], [16], understanding boundaries between reliable and chaotic operation becomes essential. Our characterization provides practitioners with principled guidance for building robust multi-agent systems.

II. Related Work

A. Numerical Stability in Deep Learning

Goldberg [6] established that floating-point operations are neither associative nor commutative, while Demmel and Nguyen [7] proved parallel reductions are inherently non-deterministic. Recent work documented these issues in deep learning: Zhuang et al. [10] showed non-reproducible results across identical seeds, Nagarajan et al. [17] quantified 40%+ variance in deep RL, and Shanmugavelu et al. [9] demonstrated non-deterministic weight evolution. Yuan et al. [18] systematically characterized numerical sources of nondeterminism in LLM inference and proposed mitigation strategies. While quantization studies [19]–[21] extensively analyzed reduced-precision arithmetic, we examine *inherent instability at full Float32 precision*.

B. Sensitivity and Chaos in Neural Networks

Hahn [11] proved Transformer condition numbers grow exponentially with sequence length. Recent work explored chaos in neural networks: Liu et al. [22] deliberately exploited chaotic dynamics for improved performance, Terada and Toyoizumi [23] showed chaos facilitates Bayesian sampling in RNNs, and Engelken et al. [24] computed Lyapunov spectra for recurrent networks. However, these focus on *beneficial* chaos or RNN-specific dynamics, not unintended chaos from floating-point errors in Transformers. For adversarial robustness, Wallace et al. [25] discovered universal triggers and Zou et al. [26] achieved 80%+ jailbreak rates, but these involve semantic-level perturbations rather than numerical instabilities.

C. LLM Robustness and Multi-Agent Systems

Kim et al. [12] showed 13.78% of correct answers flip due to persona changes, while Vijayaraj [13] found 10^{-3} magnitude perturbations cause 12.4% diagnostic flips. Tuck et al. [27] measured embedding sensitivity for adversarial detection, but focused on semantic perturbations. In multi-agent systems, Wu et al. [4] reported 23% AutoGen failure rates and Hong et al. [5] documented 31% non-reproducible MetaGPT outputs. Guo et al. [14] identified brittle communication behaviors but attributed failures to algorithmic factors rather than numerical instability.

Unlike quantization studies that examine reduced precision [19], [20] or work deliberately exploiting chaos [22], [23], we characterize **unintended chaotic behaviors from floating-point rounding at Float32**. While prior sensitivity analyses [11], [27] focus on theoretical bounds or semantic perturbations, and Yuan et al. [18] address determinism in inference pipelines, we provide the **first systematic layer-wise analysis showing machine-epsilon perturbations** ($\sim 10^{-14}$) **either amplify to around** $10^{-6}\times$ **or vanish, defining three operating regimes** (Constant, Chaotic, Signal-Dominated) with empirical validation across multiple LLM architectures.

III. Methodology

To quantify the stability of an LLM with respect to specific input perturbations, we adopt a directional approach [28]. While the standard condition number typically denotes the worst-case sensitivity (the spectral norm of the Jacobian) [29], this metric is often overly pessimistic for high-dimensional neural networks, such as LLMs.

Instead, we utilize the norm of the directional derivative,[1] which we refer to as the **absolute directional condition number**. Following the foundational theory of conditioning by Rice [30], the absolute condition number is the limit of the error magnification factor. By restricting this definition to a specific perturbation direction v, we obtain a precise measure of local stability.

Let $f : \mathbb{R}^n \to \mathbb{R}^m$ represent the LLM mapping from the embedding space to the output space. For an input $x \in \mathbb{R}^n$ and a normalized perturbation direction $v \in \mathbb{R}^n$ (where $\|v\|_2 = 1$), the absolute directional condition number $\kappa_{abs}(f, x, v)$ is defined as:

$$\kappa_{abs}(f, x, v) := \|L_f(x; v)\|_2 = \|J(x)v\|_2 \tag{1}$$

where $L_f(x; v)$ denotes the directional derivative of f at x in the direction v [31], and $J(x)$ is the Jacobian matrix. Numerically, this estimates the immediate absolute change in the output given a unit perturbation along v:

$$\kappa_{abs}(f, x, v) \approx \frac{\|f(x + \epsilon v) - f(x)\|_2}{\epsilon} \quad \text{for } \epsilon \to 0. \tag{2}$$

A value of $\kappa_{abs} \gg 1$ indicates that the model is numerically unstable in the direction v, expanding small input noises into significant output deviations.

A challenge applying κ_{abs} to LLMs is that their final outputs are probabilistic in nature, which depend on the choice of hyperparameters (such as temperature) and the sampling algorithm to be used. To overcome the issue, we use the output before the last language modeling head layer. To define it unambiguously, we decompose the LLM computation into two parts, as given

$$P(y \mid x) = \text{softmax}(\underbrace{W_U \cdot f(x)}_{\text{Logits } z}) \tag{3}$$

where:

- $f(x) \in \mathbb{R}^{d_{model}}$ denotes the final hidden state from the Transformer backbone, which is named as the last pseudo token (LST) in this paper as it is a continuous vector, not limited to the ones in the model's dictionary.

[1]Mathematically, there are two types of directional derivative: Gâteaux derivative and Fréchet derivative. The differences do not change the final solution here.

- $W_U \in \mathbb{R}^{V \times d_{model}}$ is the unembedding matrix (or output projection).
- $z = W_U \phi(x)$ represents the unnormalized logits.

As we show numerically, for LLMs, when estimated numerically, κ_{abs} can be several orders of magnitude larger than the singular value of the Jacobian, especially for small ϵ along the directions where the singular values are small, indicating that the stability of the LLMs is affected by numerical issues that have not been explored. Furthermore, we show that the numerical estimation of κ_{abs} converges to a value independent of the singular value of the direction before it vanishes. While the phenomenon is very important and has directly related to the issues reported in earlier studies such as [18], it has not been studied. We show the phenomenon is universal to LLMs.

Experimental Setting:

To rigorously validate these regions, we conducted experiments across different model architectures and datasets.

Models and Hardware:

We evaluated `Meta-Llama-3.1-8B-Instruct` and `OpenAI-GPT-OSS-20B`. To capture precise floating-point behaviors, specific hardware configurations were required. `Llama-3.1-8B` experiments were conducted on dual NVIDIA RTX A5000 GPUs (24GB VRAM each). Due to memory constraints when enforcing Float32 precision on the larger model, `GPT-OSS-20B` experiments were executed on an Intel Core i9-10900X CPU.

Datasets:

We utilized *TruthfulQA* to evaluate stability on general knowledge and reasoning tasks. Additionally, we employed *AdvBench* (subset of harmful behaviors) to test stability on adversarial prompts, determining if safety guardrails introduce additional numerical instability.

IV. Results and Analysis

A. Directional sensitivity is largely epsilon-driven

We define the **effective directional condition number** $D(\epsilon, v) := \|f(x+\epsilon v) - f(x)\|_2 / \epsilon$ as the finite-ϵ approximation to $\kappa_{abs}(f, x, v)$.

To determine whether directional sensitivity is determined by the Jacobian spectrum (as in classical local conditioning) or whether it is governed mainly by the perturbation scale ϵ in floating-point implementations, we analyze the effective directional number across multiple singular directions spanning the spectrum.

Figure 1 plots the effective directional number $D(\epsilon, v)$ at the last representative layer (`Llama-3.1-8B`) for five singular directions spanning the spectrum $(v_1, v_{50}, v_{500}, v_{2000}, v_{4096})$. The curves exhibit the same qualitative scale dependence: at sufficiently small ϵ, the response becomes dominated by representational granularity and finite-precision effects rather than by the singular value ranking. This is the sense in which the implementation's *effective* directional sensitivity can be largely ϵ-driven, even though the underlying Jacobian spectrum is well-defined.

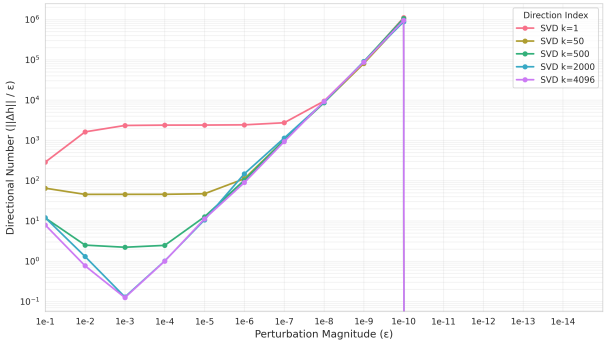


Fig. 1: Overview of effective directional number $D(\epsilon, v)$ at layer index 32 (Float32) for $v_1, v_{50}, v_{500}, v_{2000}, v_{4096}$.

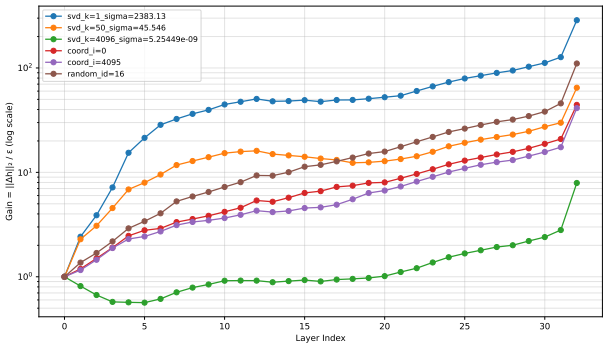


Fig. 2: Layer-wise propagation profile at $\epsilon = 0.1$. Directional structure is visible: higher-σ directions yield larger amplification than low-σ directions, consistent with a signal-dominated regime.

This observation is not a claim about the ideal mathematical condition number (defined by a limit as $\epsilon \to 0$ for a real-valued function). Instead, it characterizes the behavior of the finite-precision program under finite perturbations, where the spacing between representable numbers (ULP) depends on magnitude and can create threshold effects as ϵ changes. [6] [32]

B. Layer-wise propagation shows spectrum collapse at small epsilon

Sensitivity can either grow or vanish through network depth. If the model amplifies microscopic differences, layer-wise growth can overwhelm the initial directional structure. To separate these behaviors, we compare propagation at a large perturbation scale and a microscopic perturbation scale and identify when directional structure is preserved versus when it collapses.

Figure 2 shows the layer-wise propagation behavior at a large perturbation ($\epsilon = 0.1$). In this regime, the ordering across directions broadly follows the singular spectrum: the top singular direction exhibits larger amplification while low-σ directions remain comparatively suppressed, consistent with a signal-dominated behavior.

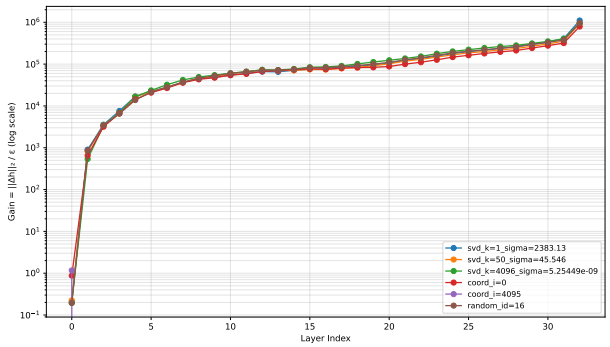


Fig. 3: Layer-wise propagation profile at $\epsilon = 10^{-10}$. Directional structure collapses: multiple singular directions, coordinate directions, and a random direction exhibit similar growth. This supports an instability-dominated regime where effective sensitivity is governed by scale rather than singular value.

Figure 3 shows the same layer-wise plot for a much smaller perturbation ($\epsilon = 10^{-10}$). Here the direction dependence largely collapses: singular directions with widely different σ_k, coordinate directions, and a random direction follow similar amplification trajectories and reach very large gains in later layers. This indicates an instability-dominated regime in which finite-precision effects and depth-wise amplification can dominate the directionality implied by the Jacobian spectrum.

Taken together, the overview and layer-wise results support an avalanche-like mechanism: once a perturbation survives rounding at the embedding interface, it can be amplified through depth in a way that becomes weakly dependent on the initial direction.

C. Microscopic instability and constant-like plateaus across prompts

We quantify how often microscopic perturbations cause noticeable representation changes and how this varies across datasets and models, revealing the prevalence of constant-like regions versus chaotic jumps.

We quantify local instability along a one-dimensional sweep in the embedding space aligned with a locally sensitive direction. For an ordered sweep $\epsilon_1 < \cdots < \epsilon_T$ and LPTs $m_i = M(x_{\epsilon_i,v})$, we define a finite-difference instability statistic

$$I_i = \frac{\|m_i - m_{i-1}\|_2}{\epsilon_i - \epsilon_{i-1}}, \quad i = 2, \ldots, T. \tag{5}$$

We report the mean and median of $\{I_i\}$, along with the maximum drift $\max_i \|m_i - m_1\|_2$ over the sweep. We also report the logit margin at the last prompt position, defined as the difference between the largest and second-largest logits; we summarize it by its mean and minimum over the sweep.

Table I shows a consistent spiky pattern under Float32 with a microscopic perturbation range: the median instability is 0.0 while the mean instability is orders of magnitude larger. This indicates that most consecutive perturbation steps produce no measurable change in the LPT (constant-like plateaus), but rare steps trigger discrete representation jumps whose finite-difference slopes become extremely large because the step size is tiny.

TABLE I: AGGREGATE STABILITY METRICS. `Llama-3.1-8B` RESULTS USE 100 PROMPTS PER DATASET (GPU). `GPT-OSS-20B` RESULTS USE 10 PROMPTS PER DATASET (CPU). MEDIAN INSTABILITY OF 0.0 INDICATES PREVALENT CONSTANT-LIKE PLATEAUS, WHILE LARGE MEAN INSTABILITY INDICATES RARE BUT LARGE FINITE-DIFFERENCE SPIKES.

Model	Dataset	Mean Inst.	Median Inst.	Max Drift	Mean Margin	Min Margin
GPT-OSS	AdvBench	6.76×10^{5}	0.0	2.92×10^{-4}	0.6196	0.6196
GPT-OSS	TruthfulQA	2.63×10^{5}	0.0	3.43×10^{-4}	0.9767	0.9767
Llama-3.1	AdvBench	3.78×10^{6}	0.0	1.02×10^{-4}	0.6175	0.6175
Llama-3.1	TruthfulQA	8.84×10^{6}	0.0	1.46×10^{-2}	0.5351	0.5035

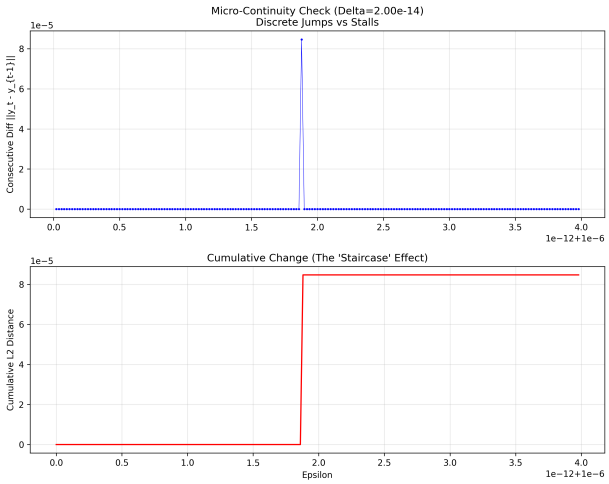


Fig. 4: Micro-continuity analysis: consecutive differences exhibit long constant-like stalls with rare discrete jumps, producing a staircase-like cumulative change in the output representation.

Micro-continuity at sub-ULP steps: To make the plateau-and-jump mechanism explicit, we probe $M(x)$ using consecutive perturbations separated by a Δ that is far smaller than typical Float32 spacing at the embedding magnitude. Figure 4 shows long runs of consecutive differences indistinguishable from zero (stalls) punctuated by discrete jumps, producing a staircase cumulative change. This explains why median instability becomes exactly zero under tiny steps while the mean can remain very large: dividing a rare jump by a tiny Δ yields a huge finite-difference slope even when the absolute change in representation is small.

D. Chaotic decision boundaries and regions

We examine whether microscopic representation changes affect downstream outputs. They matter most when the model is near a tie between the top candidate tokens (i.e., when the logit margin is small). We construct near-tie scenarios

TABLE II: DECISION BOUNDARY IRREGULARITY METRICS ACROSS SINGULAR VECTOR PLANES NEAR A NEAR-TIE POINT x_0.

Metric	v_1-v_2	v_1-v_{10}	v_1-v_{4096}
Flip Frequency (%)	16.14	16.29	15.42
Fragmentation	861	866	779
Crossing Density	51.0	52.5	49.0

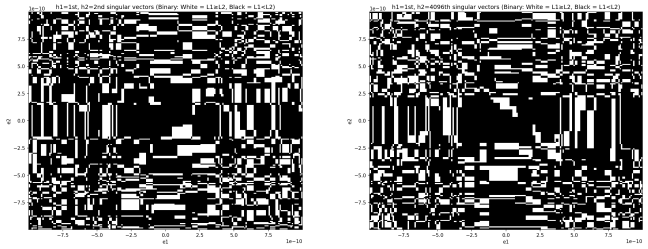
and probe two-dimensional perturbation planes to visualize decision boundary geometry.

Starting from an embedding x_0 selected so that the top two logits are nearly equal ($L_1(x_0) \approx L_2(x_0)$), we perturb in two-dimensional planes spanned by right singular vectors: $x = x_0 + \epsilon_1 v_i + \epsilon_2 v_j$ with $\epsilon_1, \epsilon_2 \in [-10^{-8}, 10^{-8}]$ and step size 10^{-10}. We test three planes: (v_1, v_2), (v_1, v_{10}), and (v_1, v_{4096}), and record which of the top two tokens wins at each grid point.

We quantify boundary irregularity using: (1) **Flip Frequency**, the fraction of adjacent grid cells with different predictions; (2) **Region Fragmentation**, the number of disconnected prediction regions; and (3) **Zero-Crossing Density**, the number of boundary crossings per line scan normalized by grid dimensions. Table II shows that all planes produce highly fragmented decision regions with dense boundary crossings. The same qualitative behavior persists even in the (v_1, v_{4096}) plane, indicating that near-tie decision sensitivity is not confined to a small subspace of "high-σ" directions. Figure 5 visualizes this fragmentation through binary decision maps showing which token wins at each point in the perturbation grid. The salt-and-pepper patterns with no smooth regions reveal that microscopic perturbations cause erratic decision flips across the entire plane, with vertical patterns indicating rounding avalanches where cascading bit-flips trigger sudden output jumps. The visualizations use the prompt *"The capital of France is"*, chosen because its singular value spectrum ($\sigma_1 = 615.3$, $\sigma_2 = 386.9$, $\sigma_{4096} \approx 3 \times 10^{-9}$) represents typical input embeddings. Under normal conditions, the top token should have a clear lead over the second; however, after perturbation, the top two tokens exhibit nearly balanced logits ($L_1 \approx L_2$), creating ideal conditions for observing decision boundary chaos.

Directional Stability Mapping: To systematically characterize instability across the embedding manifold, we measured maximum stable perturbation magnitudes using two complementary experiments: **(1)** angular sweeps in the (v_1, v_2) plane for representation-level bit-flips, and **(2)** perturbations along all 4096 singular vectors.

a) Angular Stability in the (v_1, v_2) Plane.: We parameterized perturbations as $\delta = s(\cos\theta \cdot v_1 + \sin\theta \cdot v_2)$ with $\theta \in [0, 2\pi)$ sampled at 1000 angles. For each direction, we performed exponential search followed by binary search with ULP refinement to identify $s_{\max}(\theta)$, the maximum perturbation magnitude before representation bit-flip. This ULP-precise approach ensures we find the exact floating-point boundary where $M(x_0) = M(x_0 + s_{\max}\delta)$ but $M(x_0 + s_{\text{next}}\delta) \neq M(x_0)$, where $s_{\text{next}} = \texttt{nextafter}(s_{\max})$ in the float32 lattice.



(a) v_1-v_2 plane (b) v_1-v_{4096} plane

Fig. 5: Binary decision maps near a constructed near-tie point. White: $L_1 \geq L_2$ (Token 1 wins). Black: $L_1 < L_2$ (Token 2 wins). The fragmented regions indicate that near-tie decisions can flip under microscopic perturbations, including along low-σ directions.

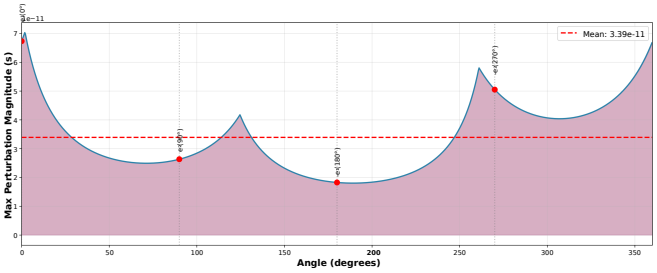
Figure 6 reveals extreme directional variability spanning four orders of magnitude. The angular profile shows $s_{\max}$ ranges from 1.80×10^{-11} at $\theta \approx 180°$ to 7.03×10^{-11} at $\theta \approx 270°$, with mean 3.39×10^{-11}. The Cartesian projection reveals an asymmetric polygonal boundary, not an ellipse as singular value theory would predict, but a distorted polygon with sharp vertices indicating catastrophic fragility in specific directions and extended lobes indicating Deep Constant Regions in others.

b) Universal Instability Across All Singular Vectors.: To test whether instability is confined to high-sensitivity subspaces, we measured $s_{\max}$ along all 4096 singular vectors v_i using identical binary search methodology. Figure 7 demonstrates that **instability is universal**: perturbations along v_1 ($\sigma_1 = 615.3$), v_{50} ($\sigma_{50} = 31.1$), and v_{4096} ($\sigma_{4096} \approx 0$) all exhibit $s_{\max} \sim 10^{-10}$, varying only by $3\times$ despite the singular values spanning from near-zero to over 600. The mean $s_{\max} = 8.96 \times 10^{-11}$ with standard deviation 4.12×10^{-11} confirms that maximum stable perturbation is determined by floating-point quantization effects rather than spectral properties. Random spikes to $s_{\max} \sim 2.5 \times 10^{-10}$ occur sporadically across the entire singular spectrum, indicating that certain directions accidentally align with rounding plateaus, these correspond to the extended lobes observed in Figure 6b.

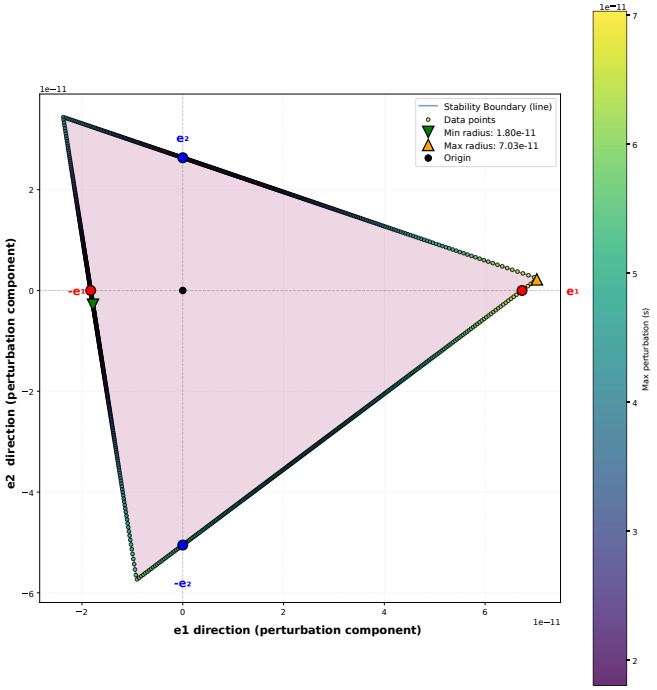
These experiments establish a critical finding: the chaotic decision boundaries are not artifacts of analyzing specific high-sensitivity planes, but reflect fundamental instability that **pervades the entire 4096-dimensional embedding space**. The fact that $s_{\max}$ varies by only $3\times$ across six orders of magnitude in singular values demonstrates that stability boundaries are determined by input-dependent rounding dynamics, not by the Jacobian's spectral structure.

E. Effects of floating-point precision

Assessing whether the three-regime picture is specific to Float32 or persists under other numeric formats shows how representational granularity affects the onset and location of instability-dominated behavior. We repeat directional sensitivity analysis under BFloat16 and Float64 precision.



(a) Angular stability profile



(b) Boundary in (v_1, v_2) space

Fig. 6: Directional stability in the (v_1, v_2) plane. (a) Maximum stable perturbation $s_{\max}(\theta)$ varies by $4\times$ across angles. (b) Cartesian projection reveals asymmetric polygonal boundary. Geometric distortion demonstrates that stability is determined by input-dependent rounding dynamics, not by singular values.

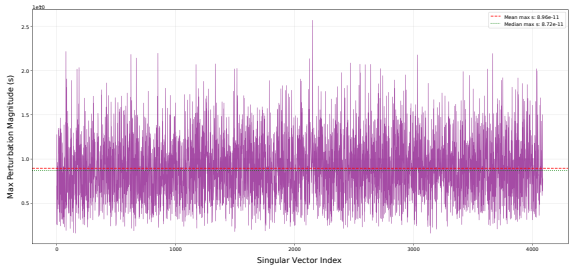
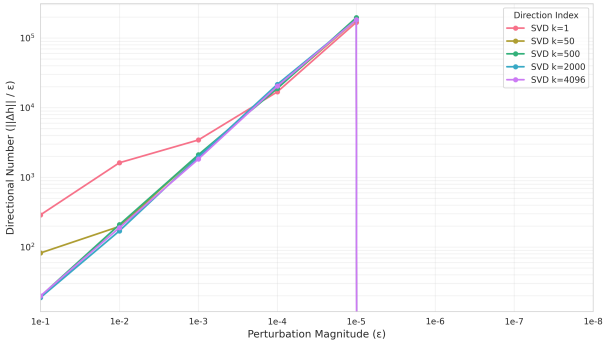
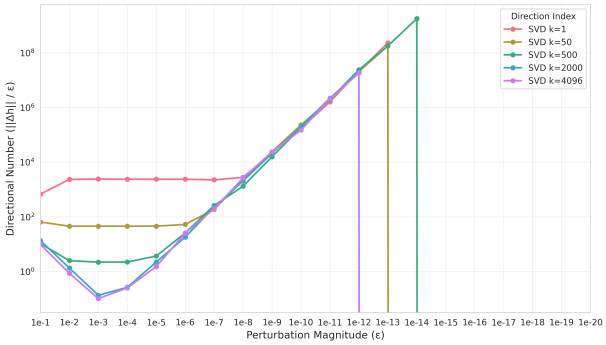


Fig. 7: Maximum stable perturbation magnitude along all 4096 singular vectors. Despite singular values spanning $\sigma_1 = 615.3$ to $\sigma_{4096} \approx 0$, the stability boundary remains nearly constant ($s_{\max} \sim 10^{-10}$) across the entire spectrum. This universality proves instability pervades the entire embedding manifold.



(a) BFloat16



(b) Float64

Fig. 8: Directional number vs. ϵ at layer index 32 under different floating-point precisions. Precision shifts the ϵ ranges where plateau behavior and rapid growth appear, while preserving the overall scale dependence.

Figure 8 repeats the directional-number sweep at the same layer index using BFloat16 and Float64. The qualitative structure remains: there is a regime where the curves partially collapse across directions and a regime where directionality becomes more visible at larger perturbations. What changes with precision is the *location* of the transition scales. In BFloat16, the onset of plateau behavior and the transition to large effective directional numbers occurs at coarser ϵ because representational spacing is larger. In Float64, the same effects are shifted toward much smaller ϵ, consistent with a finer representational grid and a wider range of distinguishable perturbations before stalling dominates.

These results support the claim that the regimes are not tied to a single floating-point type. Precision changes how quickly perturbations are rounded away and where the instability-dominated behavior becomes visible, but it does not remove the scale dependence of effective sensitivity when probing finite-ϵ behavior [6] [32].

F. Mitigation via Noise Averaging

Since the root cause of instability is the propagation of floating-point rounding errors through deep networks, we propose a simple but effective mitigation strategy: averaging multiple forward passes with injected noise. For a perturbation $x_0 + \epsilon \cdot s_i$ along singular vector s_i, instead of computing a

single absolute directional condition number estimate $\kappa_{abs} = \|f(x_0 + \epsilon s_i) - f(x_0)\|/\epsilon$, we compute a smoothed estimate by averaging n noisy evaluations: $\kappa_{\text{smooth}} = \|\frac{1}{n}\sum_{j=1}^{n} f(x_0 + \epsilon s_i + v_j) - \frac{1}{n}\sum_{j=1}^{n} f(x_0 + v_j)\|/\epsilon$, where v_j are small random perturbations along s_i with magnitude 10^{-9}. The key insight is that rounding errors are non-deterministic and vary across different computational paths, while the true model sensitivity remains constant. By the law of large numbers, averaging over multiple samples cancels out the stochastic rounding noise, revealing the underlying algorithmic sensitivity.

Figure 9 demonstrates the effectiveness of this approach. Without averaging ($n = 1$), the empirical absolute directional condition number is above 900, exceeding the theoretical maximum singular value of 615.31 due to rounding error amplification. With just $n = 10$ samples, the estimate drops sharply to ∼600, and by $n = 100$, it stabilizes close to 600, converging to the true singular value. Further increases to $n = 4000$ provide minimal additional benefit, confirming convergence. This mitigation transforms the chaotic, unreliable single-sample estimate into a stable, reproducible measurement that accurately reflects genuine model sensitivity rather than numerical artifacts. The method's practicality is demonstrated by its rapid convergence, even modest averaging ($n = 10$-100) suffices to recover reliable estimates at negligible computational cost.

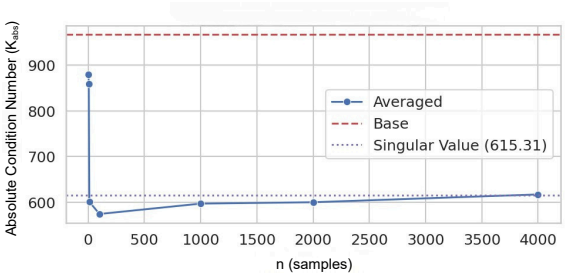


Fig. 9: Absolute directional condition number convergence via noise averaging. The dashed red line shows the unaveraged baseline, while the blue curve shows the averaged estimate stabilizing at ∼600 by $n = 100$ samples, converging to the theoretical singular value (dotted line, 615.31). Averaging eliminates rounding-error artifacts and recovers true model sensitivity.

V. Discussion

Our findings show that LLMs operate at the boundary of numerical chaos, where perturbations exhibit binary outcomes: complete attenuation in Constant Regions or explosive amplification in Chaotic Regions. Directional sensitivity is scale-driven rather than spectrum-driven, and perturbations across all singular directions (spanning five orders of magnitude in singular values) exhibit nearly identical stability thresholds, contradicting classical conditioning theory. Instability originates from early-layer avalanche effects where microscopic errors cascade through depth, producing median instability of zero but extreme mean values, confirming discrete jumps rather than smooth divergence.

Near decision boundaries, microscopic perturbations fragment output space into hundreds of disconnected regions with crossing densities exceeding smooth expectations by 50×. This fractal geometry pervades the entire 4096-dimensional embedding space. For multi-agent systems, non-deterministic floating-point operations cause identical inputs to traverse different computational paths, explaining 23-31% failure rates. Architectural differences affect stability: GPT-OSS-20B maintains larger Constant Regions and higher logit margins than Llama-3.1, suggesting normalization schemes significantly impact vulnerability. Increasing precision merely shifts regime boundaries without eliminating chaos, distributed systems cannot guarantee reproducibility due to non-associative reductions.

Limitations and Future Work. Our analysis focuses on inference-time instability. Future work should investigate: (1) whether training implicitly navigates toward stable regions; (2) architectural modifications to expand Constant Regions; (3) runtime boundary detection for adaptive precision; (4) connections between numerical chaos and adversarial vulnerability; and (5) extension to multimodal models.

VI. Conclusion

We demonstrate that LLMs exhibit universal scale-dependent chaotic behaviors arising from floating-point rounding errors amplified through transformer depth. We identify three operating regimes: Constant (bitwise-frozen outputs), Chaotic (rounding-error dominated), and Signal-Dominated (input variations override noise). Directional sensitivity is determined by perturbation scale ϵ rather than Jacobian spectrum, all singular directions exhibit $s_{\max} \sim 10^{-10}$ despite five orders of magnitude variation in singular values. Instability originates from early-layer avalanche effects, and near decision boundaries fragments output space into hundreds of chaotic regions. These findings establish numerical stability as a fundamental constraint on LLM reproducibility across heterogeneous deployments, providing practitioners with a stability framework for safety-critical applications.

References

[1] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, *et al.*, "A survey on large language model based autonomous agents," *Frontiers of Computer Science*, vol. 18, no. 6, p. 186345, 2024.

[2] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, *et al.*, "The rise and potential of large language model based agents: A survey. arxiv 2023," *arXiv preprint arXiv:2309.07864*, vol. 10, 2025.

[3] W. Chen, Y. Su, J. Zuo, C. Yang, C. Yuan, C.-M. Chan, H. Yu, Y. Lu, Y.-H. Hung, C. Qian, *et al.*, "Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors," in *The Twelfth International Conference on Learning Representations*, 2023.

[4] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. H. Awadallah, R. W. White, D. Burger, and C. Wang, "Autogen: Enabling next-gen llm applications via multi-agent conversation," 2023.

[5] S. Hong, M. Zheng, J. Chen, X. Cheng, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin, L. Zhou, C. Ran, *et al.*, "Metagpt: Meta programming for a multi-agent collaborative framework," in *The Twelfth International Conference on Learning Representations*, 2024.

[6] D. Goldberg, "What every computer scientist should know about floating-point arithmetic," *ACM computing surveys (CSUR)*, vol. 23, no. 1, pp. 5–48, 1991.

[7] J. Demmel and H. D. Nguyen, "Parallel reproducible summation," *IEEE Transactions on Computers*, vol. 64, no. 7, pp. 2060–2070, 2014.

[8] NVIDIA Corporation, "cuDNN developer guide." https://docs.nvidia.com/deeplearning/cudnn/, 2024.

[9] S. Shanmugavelu, M. Taillefumier, C. Culver, O. Hernandez, M. Coletti, and A. Sedova, "Impacts of floating-point non-associativity on reproducibility for hpc and deep learning applications," in *SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 170–179, IEEE, 2024.

[10] D. Zhuang, X. Zhang, S. Song, and S. Hooker, "Randomness in neural network training: Characterizing the impact of tooling," in *Proceedings of Machine Learning and Systems* (D. Marculescu, Y. Chi, and C. Wu, eds.), vol. 4, pp. 316–336, 2022.

[11] M. Hahn, "Theoretical limitations of self-attention in neural sequence models," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 156–171, 2020.

[12] J. Kim, N. Yang, and K. Jung, "Persona is a double-edged sword: Mitigating the negative impact of role-playing prompts in zero-shot reasoning tasks," 2024.

[13] R. K. Vijayaraj, "Embeddings to diagnosis: Latent fragility under agentic perturbations in clinical llms," 2025.

[14] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, and X. Zhang, "Large language model based multi-agents: A survey of progress and challenges," *arXiv preprint arXiv:2402.01680*, 2024.

[15] PyTorch Contributors, "Reproducibility in PyTorch." https://pytorch.org/docs/stable/notes/randomness.html, 2024.

[16] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[17] P. Nagarajan, G. Warnell, and P. Stone, "Deterministic implementations for reproducibility in deep reinforcement learning," 2019.

[18] J. Yuan, H. Li, X. Ding, W. Xie, Y.-J. Li, W. Zhao, K. Wan, J. Shi, X. Hu, and Z. Liu, "Understanding and mitigating numerical sources of nondeterminism in LLM inference," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[19] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *Advances in neural information processing systems*, vol. 36, pp. 10088–10115, 2023.

[20] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, "Gptq: Accurate post-training quantization for generative pre-trained transformers," *arXiv preprint arXiv:2210.17323*, 2022.

[21] G. Xiao, J. Lin, M. Seznec, H. Wu, J. Demouth, and S. Han, "Smoothquant: Accurate and efficient post-training quantization for large language models," in *International conference on machine learning*, pp. 38087–38099, PMLR, 2023.

[22] S. Liu, N. Akashi, Q. Huang, Y. Kuniyoshi, and K. Nakajima, "Exploiting chaotic dynamics as deep neural networks," *Physical Review Research*, vol. 7, no. 3, p. 033031, 2025.

[23] Y. Terada and T. Toyoizumi, "Chaotic neural dynamics facilitate probabilistic computations through sampling," *Proceedings of the National Academy of Sciences*, vol. 121, no. 18, p. e2312992121, 2024.

[24] R. Engelken, F. Wolf, and L. F. Abbott, "Lyapunov spectra of chaotic recurrent neural networks," *Physical Review Research*, vol. 5, no. 4, p. 043044, 2023.

[25] E. Wallace, S. Feng, N. Kandpal, M. Gardner, and S. Singh, "Universal adversarial triggers for attacking and analyzing nlp," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2153–2162, 2019.

[26] A. Zou, Z. Wang, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," *arXiv preprint arXiv:2307.15043*, 2023.

[27] B. E. Tuck and R. M. Verma, "Assessing representation stability for transformer models," *arXiv preprint arXiv:2508.11667*, 2025.

[28] Y. Nesterov, "Efficiency of coordinate descent methods on huge-scale optimization problems," *SIAM Journal on Optimization*, vol. 22, no. 2, pp. 341–362, 2012.

[29] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*. Philadelphia, PA: SIAM, 2nd ed., 2002.

[30] J. R. Rice, "A theory of condition," *SIAM Journal on Numerical Analysis*, vol. 3, no. 2, pp. 287–310, 1966.

[31] N. J. Higham, *Functions of Matrices: Theory and Computation*. Philadelphia, PA: SIAM, 2008. See Chapter 3 for Fréchet derivatives.

[32] IEEE, "Ieee standard for floating-point arithmetic," *IEEE Std 754-2019 (Revision of IEEE 754-2008)*, pp. 1–84, 2019.