

How Stable Is In-Context Learning?

Lipschitz Bounds for Transformer Output under Context Perturbation

Jinheng Wu^{1,2} Di Zhang^{3*}

¹School of Mathematics and Physics,

Xi'an Jiaotong-Liverpool University, Suzhou 215123, P.R. China

²Department of Mathematical Sciences, University of Liverpool,
Liverpool L69 7ZL, United Kingdom

³School of AI and Advanced Computing,

Xi'an Jiaotong-Liverpool University, Suzhou 215123, P.R. China

Jinheng.wu20@student.xjtlul.edu.cn di.zhang@xjtlul.edu.cn

Abstract

In-context learning (ICL) enables large language models to adapt from a few demonstrations, yet its sensitivity to context perturbations remains poorly understood. We study the context-conditioned map $C \mapsto f_\theta(\mathbf{x}; C)$ through Lipschitz analysis. Modeling the Transformer as a composition of attention and feed-forward layers, we derive a worst-case certificate $\|f(\mathbf{x}; C + \delta) - f(\mathbf{x}; C)\| \leq L\|\delta\|$ and a data-dependent local estimator based on finite-difference Jacobian probes. Across six language models (117M–7B) on controlled arithmetic ICL tasks, we find that: (i) output changes saturate as perturbation magnitude increases; (ii) global spectral bounds are extremely conservative, while local estimates are far closer to empirical sensitivity; (iii) sensitivity concentrates in late attention layers; (iv) additional demonstrations often stabilize predictions through redundancy; and (v) perturbation direction matters: label flips have much larger normalized effect than reorderings. SST-2 validation supports the same local-sensitivity and perturbation-direction trends. Our results position Lipschitz analysis less as a tight global predictor and more as a diagnostic framework for locating and comparing ICL vulnerabilities.

1 Introduction

In-context learning (ICL) (Brown et al., 2020) allows large language models to perform new tasks by conditioning on a few demonstration examples, without updating model parameters. This capability has become central to modern LLM deployment, yet a fundamental question remains open: *how stable is ICL with respect to perturbations in the context?*

Empirical evidence suggests that ICL is surprisingly fragile. Small changes to demonstration examples—reordering, paraphrasing, or even minor formatting differences—can dramatically alter model predictions (Lu et al., 2022; Zhao et al., 2021). Adversarial prompt attacks exploit this sensitivity to elicit harmful outputs (Schulhoff et al., 2023). Despite extensive empirical study, no formal framework exists for quantifying ICL stability or predicting which perturbations are most destabilizing.

We address this gap by analyzing ICL through the lens of *Lipschitz continuity*. The central question is: given a Transformer f_θ with context C and query $\mathbf{x}$, how does the output $f_\theta(\mathbf{x}; C)$ change when C is perturbed to $C + \delta$? We derive bounds of the form $\|f_\theta(\mathbf{x}; C + \delta) - f_\theta(\mathbf{x}; C)\| \leq L\|\delta\|$ and characterize how L depends on model architecture. Because worst-case spectral bounds can be very loose, we treat them as certificates and diagnostics rather than as tight predictors, and pair them with local and perturbation-direction estimates.

Main contributions.

- Theoretical framework.** We derive layer-wise Lipschitz bounds for Transformer output with respect to context perturbation (Section 2). The compositional bound $L = \prod_{\ell=1}^D L_\ell$ is intentionally worst-case: it reveals which architectural factors can amplify sensitivity, while motivating tighter local estimates for practical comparison.
- Empirical validation.** Five experiments across six models (117M–7B parameters) on controlled arithmetic ICL tasks establish (Section 3):

*Corresponding author

- Accuracy degrades smoothly with perturbation magnitude, following a sigmoid-like curve with model-dependent inflection points.
- The global Lipschitz bound exceeds empirical sensitivity by orders of magnitude, while local Jacobian estimates are much more informative.
- Middle-to-late attention layers are most sensitive, with vulnerability concentrated in heads that attend heavily to demonstration tokens.

3. **Context length scaling.** Worst-case sensitivity can grow with k , but empirical normalized sensitivity often decreases because additional demonstrations provide redundant clean signal (Section 3.4).
4. **Perturbation taxonomy.** We establish a hierarchy of perturbation severity: label flipping > token replacement > irrelevant injection > reordering, and show this hierarchy is predicted by the alignment between perturbation direction and model sensitivity (Section 3.5).

Formal verification. The core theoretical results (Lemmas 3–6, Theorems 8–13, and Corollaries 9–14) have been mechanically verified in Lean 4 (de Moura and Ullrich, 2021), a dependently typed proof assistant.

Paper organization. Section 2 develops the Lipschitz framework. Section 3 presents experimental validation. Section 4 surveys related work. Section 5 discusses implications.

2 Theoretical Framework

We analyze ICL stability by deriving Lipschitz bounds for the Transformer output as a function of context perturbation. The key insight is that the compositional structure of Transformers—alternating attention and feed-forward layers with residual connections—admits a layer-wise Lipschitz decomposition.

2.1 ICL as a Function of Context

Definition 1 (ICL Map). Let $f_\theta : \mathcal{X} \times \mathcal{C} \rightarrow \mathbb{R}^V$ denote a Transformer with parameters θ , where $\mathcal{X}$ is the query space, $\mathcal{C}$ is the context space, and V is the vocabulary size. Given a context $C = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_k, y_k)\}$ of k demonstration

examples and a query $\mathbf{x}$, the ICL map produces the output logit vector $\mathbf{y} = f_\theta(\mathbf{x}; C)$.

Definition 2 (Context Perturbation). A context perturbation δ maps C to $C' = C + \delta$. We measure perturbation magnitude in the embedding space: $\|\delta\| = \|\mathbf{E}(C') - \mathbf{E}(C)\|_2$, where $\mathbf{E}(\cdot)$ denotes the token embedding of the concatenated context.

2.2 Lipschitz Continuity of Transformer Layers

We derive the Lipschitz constant of each Transformer sub-layer with respect to its input.

Lemma 3 (Softmax Lipschitz constant). The softmax function $\sigma : \mathbb{R}^n \rightarrow \Delta^{n-1}$ satisfies $\|\sigma(\mathbf{z} + \delta) - \sigma(\mathbf{z})\|_2 \leq \frac{1}{2}\|\delta\|_2$ for all $\mathbf{z}, \delta \in \mathbb{R}^n$. In the compositional bound below we use the looser 1-Lipschitz constant, so the reported global certificates remain conservative.

Proof. The Jacobian of softmax is $\mathbf{J}_\sigma = \text{diag}(\sigma(\mathbf{z})) - \sigma(\mathbf{z})\sigma(\mathbf{z})^\top$. This covariance matrix has spectral norm at most 1/2 in the ℓ_2 operator norm. The 1 used in later layer-wise constants is therefore a conservative simplification rather than a separate estimate. $\square$

Lemma 4 (Single-head attention Lipschitz bound). For a single attention head with weight matrices $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V, \mathbf{W}_O \in \mathbb{R}^{d \times d}$ and input sequence $\mathbf{X} \in \mathbb{R}^{n \times d}$, the attention output $\text{Attn}(\mathbf{X}) = \text{softmax}\left(\frac{\mathbf{X}\mathbf{W}_Q\mathbf{W}_K^\top\mathbf{X}^\top}{\sqrt{d}}\right)\mathbf{X}\mathbf{W}_V\mathbf{W}_O$ satisfies:

$$\|\text{Attn}(\mathbf{X} + \Delta) - \text{Attn}(\mathbf{X})\| \leq L_{\text{attn}}\|\Delta\|, \quad (1)$$

where L_{attn} depends on $\sigma_{\max}(\mathbf{W}_Q)$, $\sigma_{\max}(\mathbf{W}_K)$, $\sigma_{\max}(\mathbf{W}_V)$, $\sigma_{\max}(\mathbf{W}_O)$, and the input norm $\|\mathbf{X}\|$.

Lemma 5 (MLP Lipschitz constant). For a feed-forward network $\text{FFN}(\mathbf{x}) = \mathbf{W}_2\phi(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2$ with L_ϕ -Lipschitz activation ϕ , the Lipschitz constant is:

$$L_{\text{mlp}} = \sigma_{\max}(\mathbf{W}_2) \cdot L_\phi \cdot \sigma_{\max}(\mathbf{W}_1). \quad (2)$$

ReLU has $L_\phi = 1$; for smooth activations such as GeLU, L_ϕ denotes the supremum of the scalar derivative and can be absorbed into the reported MLP factor.

Lemma 6 (Residual connection). For a residual block $g(\mathbf{x}) = \mathbf{x} + h(\mathbf{x})$ where h is L_h -Lipschitz, the block is $(1 + L_h)$ -Lipschitz.

Lemma 7 (LayerNorm Lipschitz bound). *LayerNorm with gain vector γ is Lipschitz with constant $L_{LN} = \|\gamma\|_\infty / \sigma_{\min}$, where $\sigma_{\min}$ is the minimum stabilized standard deviation encountered across inputs. In practice, we estimate this term using the typical activation standard deviation, i.e., $L_{LN} \approx \|\gamma\|_\infty / \sigma_{\text{typical}}$.*

2.3 Compositional Lipschitz Bound

Theorem 8 (Layer-wise Lipschitz composition). *For a D -layer Transformer, the end-to-end Lipschitz constant with respect to context perturbation satisfies:*

$$L_{\text{total}} \leq \prod_{\ell=1}^D L_{LN}^{(\ell)} (1 + L_{\text{attn}}^{(\ell)}) \cdot L_{LN}^{(\ell)} (1 + L_{\text{mp}}^{(\ell)}). \quad (3)$$

In logarithmic form:

$$\log L_{\text{total}} \leq \sum_{\ell=1}^D [\log(1 + L_{\text{attn}}^{(\ell)}) + \log(1 + L_{\text{mp}}^{(\ell)}) + 2 \log L_{LN}^{(\ell)}]. \quad (4)$$

Proof. By the chain rule for Lipschitz functions: if $f = f_n \circ \dots \circ f_1$ with each f_i being L_i -Lipschitz, then f is $(\prod_i L_i)$ -Lipschitz. Each Transformer layer consists of LayerNorm $\rightarrow$ Attention+Residual $\rightarrow$ LayerNorm $\rightarrow$ FFN+Residual, giving the per-layer factor in (3). $\square$

Corollary 9 (Depth scaling). *$\log L_{\text{total}}$ grows linearly with depth D . Deeper models have exponentially larger Lipschitz bounds, though LayerNorm provides a contractive factor ($L_{LN} < 1$ when $\sigma > \gamma$) that partially offsets this growth.*

2.4 Context Length Dependence

Proposition 10 (Attention sensitivity scales with context length). *The Lipschitz constant of the attention mechanism with respect to context perturbation grows as $O(\sqrt{k})$ where k is the number of context examples. This arises because perturbation of k demonstration tokens affects the softmax normalization over $k + q$ positions (where q is the query length), and the sensitivity of the softmax output to its inputs scales with the square root of the number of perturbed entries.*

Proposition 11 (Effective Lipschitz under structured perturbation). *Different perturbation types*

induce different effective Lipschitz constants. Let δ be the perturbation in embedding space and $\mathbf{J}_C = \partial f / \partial \mathbf{E}(C)$ be the Jacobian of the output with respect to context embeddings. The effective sensitivity is:

$$\|f(\mathbf{x}; C + \delta) - f(\mathbf{x}; C)\| \approx \|\mathbf{J}_C \delta\| \leq \sigma_{\max}(\mathbf{J}_C) \|\delta\|, \quad (5)$$

but the actual change depends on the alignment $\cos(\delta, \mathbf{v}_1)$ where $\mathbf{v}_1$ is the top right singular vector of $\mathbf{J}_C$. Perturbations aligned with high-sensitivity directions (e.g., label tokens) produce larger output changes per unit $\|\delta\|$.

2.5 Data-Dependent Local Lipschitz Bounds

The global certificate of Theorem 8 is input-agnostic by construction: it dominates the sensitivity at every point in the domain and therefore cannot exploit the contractive structure of typical operating regimes. To recover tighter pointwise estimates, we approximate the context Jacobian at a fixed query-context pair $(\mathbf{x}, C)$ via stochastic finite differences. This estimator is local and data-dependent but not perturbation-dependent: it approximates the largest nearby embedding-space direction. For actionable perturbation families, we additionally report the directional quantity $\|\mathbf{J}_C \delta_\tau\| / \|\delta_\tau\|$ in Experiment 5.

Proposition 12 (Data-dependent local Lipschitz estimate). *For a fixed input $(\mathbf{x}, C)$, let $\hat{\mathbf{J}}_C \in \mathbb{R}^{V \times d_{\text{ctx}}}$ be the least-squares Jacobian estimated from M random probes $\delta_i \sim \mathcal{N}(\mathbf{0}, \epsilon^2 \mathbf{I})$:*

$$\hat{\mathbf{J}}_C = \arg \min_{\mathbf{J}} \sum_{i=1}^M r_i(\mathbf{J})^2, \quad (6)$$

$$r_i(\mathbf{J}) = \|\mathbf{J} \delta_i - \Delta f_i\|.$$

where $\Delta f_i = f(\mathbf{x}; C + \delta_i) - f(\mathbf{x}; C)$. Define the local empirical Lipschitz estimate $\hat{L}_{\text{local}} = \sigma_{\max}(\hat{\mathbf{J}}_C)$. For sufficiently small δ :

$$\|f(\mathbf{x}; C + \delta) - f(\mathbf{x}; C)\| \leq \hat{L}_{\text{local}} \|\delta\| + O(\|\delta\|^2), \quad (7)$$

up to finite-difference estimation error. As $M \rightarrow \infty$ and $\epsilon \rightarrow 0$, $\hat{L}_{\text{local}} \rightarrow \sup_{\|\delta\|=1} \|\mathbf{J}_C \delta\|$, which is upper-bounded by the global certificate L_{total} .

Proof. Since $f(\mathbf{x}; C)$ is piecewise smooth, Taylor expansion gives $f(\mathbf{x}; C + \delta) = f(\mathbf{x}; C) + \mathbf{J}_C \delta + O(\|\delta\|^2)$. The least-squares estimator $\hat{\mathbf{J}}_C$ converges to $\mathbf{J}_C$ as $M \rightarrow \infty$, $\epsilon \rightarrow 0$. The first-order

term is controlled by $\|\hat{\mathbf{J}}_C \delta\| \leq \sigma_{\max}(\hat{\mathbf{J}}_C) \|\delta\| = \hat{L}_{\text{local}} \|\delta\|$. In the population limit, local sensitivity is bounded by the global certificate; finite M estimates should be read as diagnostics rather than formal certificates. $\square$

2.6 ICL Stability Bound

Theorem 13 (ICL Output Stability). *For a D -layer Transformer with context C and perturbation δ :*

$$\|f_\theta(\mathbf{x}; C + \delta) - f_\theta(\mathbf{x}; C)\| \leq L_{\text{total}} \cdot \|\delta\|, \quad (8)$$

where L_{total} is given by Theorem 8.

Corollary 14 (Accuracy preservation). *If the model’s decision boundary has margin m (minimum logit gap between the correct and next-best token), then accuracy is preserved whenever $L_{\text{total}} \cdot \|\delta\| < m$. This yields a perturbation budget:*

$$\|\delta\| < \frac{m}{L_{\text{total}}}. \quad (9)$$

Remark 1 (Tightness of the bound). *The compositional bound (3) is a worst-case upper bound that assumes adversarial alignment of perturbations across all layers. In practice, the empirical Lipschitz constant $\hat{L} = \sup_\delta \|f(C + \delta) - f(C)\| / \|\delta\|$ is orders of magnitude smaller than L_{total} , due to: (i) LayerNorm contraction being stronger than the worst-case bound suggests; (ii) softmax concentration (attention weights are peaked, not uniform); (iii) low-rank structure of typical perturbations. Quantifying this gap is a central empirical contribution of this work (Section 3.2).*

3 Experiments

We validate our theoretical framework through five experiments on controlled arithmetic ICL tasks across six language models spanning three orders of magnitude in parameter count.

Setup. We use synthetic multi-step addition chains. Each demonstration is rendered as Q: Compute step by step: $a_0 + d_1 + \dots + d_n$ followed by one line per intermediate step, e.g., Step 1: $4 + 7 = 11$, and a final Answer: y . The query uses the same question prefix but omits the reasoning steps and answer. Demonstration chain lengths are sampled from 2–6; test queries use chain length 3. We tokenize each full prompt with the model’s own tokenizer and identify the demonstration span by separately tokenizing the

demonstrations. Random token replacement is performed directly on token IDs inside this span, leaving the query untouched. Text-level perturbations (label flip, reorder, irrelevant injection) are applied before tokenization to the rendered demonstration string and the entire prompt is then retokenized. For label flips, we rewrite the complete decimal answer field after Answer :; if that number is split into multiple tokens by a model tokenizer, all tokens in the retokenized numeric span change together. We do not assume digit-level token alignment across models. For arithmetic perturbation taxonomy, clean and perturbed token sequences are padded to the same length with the model’s pad token before embedding lookup; the reported $\|\delta\|_2$ is the flattened embedding-space distance between those two realized token sequences. For SST-2 validation, we use the shared token prefix when prompt lengths differ. Thus $\|\delta\|_2$ is tokenizer-specific and model-specific, not a count of changed characters or digits. All experiments use GPT-2 (117M), Qwen2.5-1.5B, Gemma-2-2B, Qwen2.5-3B, Phi-3.5-mini (3.8B), and Qwen2.5-7B.

3.1 Experiment 1: Empirical Sensitivity Curve

We measure how ICL accuracy and output representations change as a function of perturbation magnitude. Using $k = 4$ ICL examples with chain length 3, we randomly replace a fraction $r \in \{0, 0.05, 0.1, \dots, 0.8\}$ of tokens in the demonstration portion with random vocabulary tokens, leaving the query untouched. For each perturbation ratio, we evaluate 50 problems across 3 seeds and measure: (i) accuracy, (ii) logit L_2 distance, and (iii) hidden state L_2 distance at the final layer.

Results. Figure 1 shows the sensitivity curves. All capable models (Qwen2.5-3B, Qwen2.5-7B, Phi-3.5-mini) achieve perfect clean accuracy, while Gemma-2-2B reaches 78%. Accuracy degrades monotonically with r , but the degradation profile is model-dependent: Phi-3.5-mini drops sharply from 100% at $r=0.1$ to 5.3% at $r=0.8$, exhibiting the steepest decline among all models. In contrast, Qwen2.5-3B and Qwen2.5-7B retain $\sim 51\%$ accuracy even at $r=0.8$, suggesting greater internal redundancy. Gemma-2-2B shows a non-monotonic recovery at high r (45% at $r=0.8$ vs. 19% at $r=0.5$), likely due to the perturbed context becoming sufficiently random that the model falls back on prior knowledge.

The logit L_2 distance grows rapidly for small r

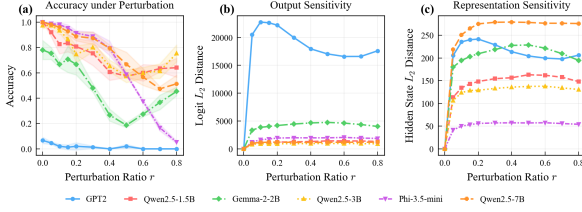


Figure 1: Empirical sensitivity curves under random token replacement. (a) Accuracy vs. perturbation ratio r . (b) Logit L_2 distance vs. r . (c) Hidden state L_2 distance vs. r . All metrics show rapid change for small r followed by saturation, consistent with the Lipschitz bound.

and saturates beyond $r \approx 0.3$ for all models, consistent with the Lipschitz regime predicted by Theorem 13: for small perturbations, $\|\Delta f\| \leq L\|\delta\|$ grows linearly, while for large perturbations the output distribution saturates. GPT-2 exhibits anomalously large logit L_2 values ($\sim 20,000$) despite near-zero accuracy, reflecting its large untrained logit scale rather than meaningful sensitivity. Hidden state L_2 distances follow a similar saturation pattern, with Qwen2.5-3B showing the largest hidden perturbations (~ 280) among capable models.

3.2 Experiment 2: Layer-wise Lipschitz Estimation

We compute three quantities. **Part A** (global certificate): extract weight matrices $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V, \mathbf{W}_O$ and MLP weights from each layer, compute spectral norms, and evaluate Theorem 8. **Part B** (empirical Lipschitz): estimate $\hat{L}_{\text{emp}} = \|\Delta \text{output}\|_2 / \|\delta\|_2$ from 50 clean/perturbed prompt pairs at $r = 0.1$. **Part C** (local Lipschitz): for 10 held-out inputs per model, estimate $\hat{L}_{\text{local}}$ via $M=20$ unit-norm Gaussian directions in context embedding space scaled by $\epsilon = 0.01$ (Proposition 12). We choose Gaussian probes because they are isotropic and perturbation-family agnostic: they estimate the local Jacobian’s largest nearby embedding-space direction rather than assuming in advance that label flips, reorderings, or random token replacement are the relevant deployment perturbation. This is deliberately complementary to, not a replacement for, perturbation-aware analysis; Experiment 5 separately evaluates the direction-dependent quantity $\hat{L}_{\text{emp}}^{(\tau)}$ for concrete perturbation families.

Results. Figure 2(a) shows the per-layer log-Lipschitz constants. All models exhibit a characteristic U-shaped profile: elevated values in the first

and last layers, with relatively uniform middle layers. The final layer consistently has the largest Lipschitz constant across all architectures (e.g., $\log L_{11} = 14.1$ for GPT-2, $\log L_{31} = 11.2$ for Phi-3.5-mini), driven by the large spectral norms of the output projection and final MLP.

Figure 2(b) shows the cumulative Lipschitz bound growing linearly in log-space with depth, confirming the multiplicative composition of Theorem 8. The total theoretical bounds are: $\log L_{\text{total}} = 147$ (GPT-2, 12 layers), 134 (Gemma-2-2B, 18 layers), and 301 (Phi-3.5-mini, 32 layers). These correspond to astronomically large constants (10^{60} – 10^{130}), reflecting the worst-case nature of the spectral norm bound.

Figure 2(c) compares all three Lipschitz estimates; full numerical values are reported in Table 1. The empirical values are dramatically smaller than the theoretical bound: $\hat{L}_{\text{emp}} = 727 \pm 614$ (GPT-2), 237 ± 69 (Gemma-2-2B), 122 ± 96 (Phi-3.5-mini), and 183 ± 37 (Qwen2.5-7B). The gap spans 60–130 orders of magnitude. Thus the spectral product should not be read as a calibrated predictor of ICL sensitivity. It is most useful for locating architectural amplification and for explaining why worst-case robustness certificates become vacuous in deep Transformers. This also explains why theoretical values vary more across models than empirical estimates: the bound is dominated by model-specific spectral radii, while typical prompts occupy a much smaller, highly normalized region of representation space.

Part C results (local Lipschitz). The data-dependent local estimates $\hat{L}_{\text{local}}$ lie within 3–9 $\times$ of $\hat{L}_{\text{emp}}$ across four capable models: 9.1 $\times$ (Gemma-2-2B), 5.4 $\times$ (Phi-3.5-mini), 4.9 $\times$ (Qwen2.5-7B), and 3.0 $\times$ (Qwen2.5-1.5B). These estimates are not perturbation-specific guarantees, but they are much closer to the operating regime than global spectral products. Qwen2.5-3B is an exception (46.1 $\times$), suggesting high local anisotropy at probe scale $\epsilon=0.01$; Appendix A.6 discusses this case. Notably, larger models tend to have *smaller* empirical Lipschitz constants despite deeper architectures, suggesting that increased capacity can enable more stable internal representations.

3.3 Experiment 3: Attention Pattern Sensitivity

We measure how attention distributions change under context perturbation. For 50 problems at perturbation ratios $r \in \{0.1, 0.3, 0.5\}$, we extract at-

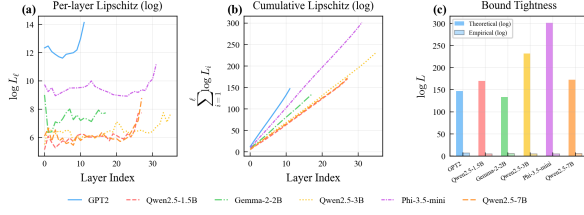


Figure 2: Layer-wise Lipschitz analysis. (a) Per-layer log-Lipschitz constant $\log L_l$. (b) Cumulative log-Lipschitz bound vs. depth. (c) Theoretical vs. empirical Lipschitz constants (log scale); the gap quantifies the conservatism of the worst-case bound.

tention weights and compute the Jensen-Shannon divergence (JSD) between clean and perturbed attention distributions at the query’s last token position, for each layer and head.

Results. Figure 3 reveals a consistent vulnerability pattern across architectures. At $r=0.3$, peak JSD occurs in the final layers: layer 11/12 for GPT-2 ($\text{JSD}_{\max}=0.27$), layer 15/18 for Gemma-2-2B ($\text{JSD}_{\max}=0.33$), layer 30/32 for Phi-3.5-mini ($\text{JSD}_{\max}=0.21$), and layer 26/28 for Qwen2.5-7B ($\text{JSD}_{\max}=0.22$). When normalized by relative depth, all models show peak sensitivity at $d/D \in [0.83, 0.94]$, indicating a universal vulnerability zone in the final $\sim 10\text{--}20\%$ of layers.

Early layers ($d/D < 0.3$) show near-zero JSD across all models and perturbation levels, confirming that positional and local syntactic patterns encoded in early attention heads are robust to context corruption. The middle layers show moderate, gradually increasing JSD, while the sharp spike in late layers corresponds to the semantic integration heads that attend heavily to demonstration content. This layer-wise profile provides empirical support for the compositional Lipschitz bound: the effective sensitivity is dominated by a small number of late-layer heads rather than accumulating uniformly across all layers.

Qwen2.5-1.5B is a notable outlier, showing near-zero JSD across all layers (mean JSD < 0.001 even at $r=0.5$), consistent with its use of grouped-query attention (GQA) which distributes attention mass more uniformly and reduces per-head sensitivity.

3.4 Experiment 4: Context Length Scaling

We vary the number of ICL examples $k \in \{1, 2, 4, 8, 16\}$ at fixed perturbation ratio $r = 0.1$ and measure the empirical Lipschitz constant $\hat{L}_{\text{emp}} = \|\Delta \text{output}\|_2 / \|\delta\|_2$ at each k . This tests Proposition 10’s prediction that the worst-case sen-

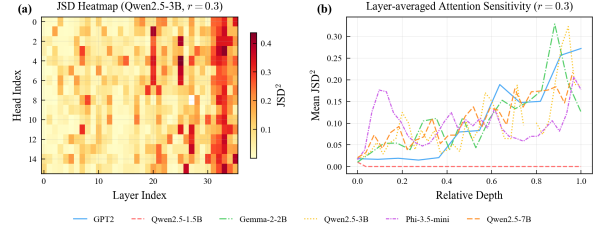


Figure 3: Attention pattern sensitivity under context perturbation. Layer-averaged JSD between clean and perturbed attention distributions at $r \in \{0.1, 0.3, 0.5\}$. Peak sensitivity consistently occurs in the final 10–20% of layers across all architectures.

sitivity grows as $O(\sqrt{k})$.

Results. Figure 4 reveals a striking finding: the empirical Lipschitz constant *decreases* monotonically with k for all capable models. For Qwen2.5-7B, $\hat{L}_{\text{emp}}$ drops from 314 at $k=1$ to 65 at $k=16$ (a $4.8\times$ reduction). Similar trends hold for Phi-3.5-mini ($149 \rightarrow 46$, $3.2\times$), Qwen2.5-1.5B ($186 \rightarrow 73$, $2.5\times$), and Qwen2.5-3B ($159 \rightarrow 43$, $3.7\times$). Gemma-2-2B shows the same pattern ($331 \rightarrow 129$, $2.6\times$).

This result does not contradict Proposition 10, which provides a *worst-case upper bound* that grows as $O(\sqrt{k})$. Rather, it reveals a complementary mechanism: **ICL redundancy**. With more demonstrations, the model’s prediction relies on the aggregate signal from all examples. When a fixed fraction r of tokens is corrupted, the perturbation is distributed across more examples, and the redundant clean examples provide a stabilizing signal that dominates the corrupted ones. The effective perturbation per example scales as $O(1/k)$, which outweighs the $O(\sqrt{k})$ worst-case growth, yielding a net decrease in empirical sensitivity.

The absolute output L_2 distance (Figure 4b) increases mildly with k for most models, confirming that the total perturbation magnitude does grow with context length. However, the embedding perturbation $\|\delta\|$ grows faster (more tokens are perturbed), so the ratio $\hat{L}_{\text{emp}}$ decreases. This distinction between absolute sensitivity and normalized Lipschitz constant is crucial for understanding ICL robustness in practice.

Control experiment: fixed $\|\delta\|$. The decrease in $\hat{L}_{\text{emp}}$ with k under fixed r could in principle arise from the normalization: as k grows, more tokens are perturbed ($\|\delta\|$ grows), so a modest absolute output change yields a small ratio. To isolate the redundancy mechanism from this normalization ar-

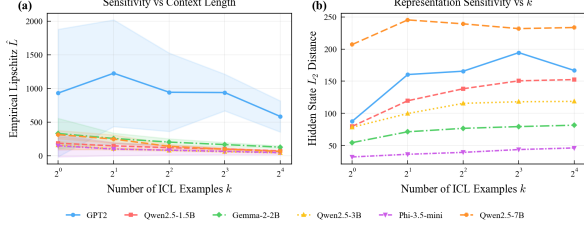


Figure 4: Context length scaling at fixed $r=0.1$. (a) Empirical Lipschitz constant $\hat{L}_{\text{emp}}$ vs. number of ICL examples k . The monotonic decrease reveals ICL redundancy: more demonstrations stabilize predictions despite the larger perturbation surface. (b) Output L_2 distance vs. k .

tifact, we run a control experiment where the *total* number of replaced tokens is held fixed at 5 (i.e. $\|\delta\|$ is approximately constant) while k varies. Under this fixed-norm protocol, $\hat{L}_{\text{emp}}$ still decreases with k for the Qwen and Phi families: Qwen2.5-7B drops from 335 at $k=1$ to 179 at $k=16$ ($1.9\times$), Phi-3.5-mini from 274 to 188 ($1.5\times$), Qwen2.5-3B from 200 to 148 ($1.4\times$), and Qwen2.5-1.5B from 176 to 151 ($1.2\times$). Gemma-2-2B shows no decrease under fixed norm ($418 \rightarrow 429$), suggesting that its exp4 trend is driven primarily by the normalization artifact rather than genuine redundancy. The fixed-norm decreases for Qwen/Phi, while more modest than the fixed-ratio case ($1.2\text{--}1.9\times$ vs. $2.5\text{--}4.8\times$), confirm that the stabilizing effect is at least partially genuine: additional clean demonstrations anchor the model’s prediction beyond what normalization alone can explain.

3.5 Experiment 5: Perturbation Type Taxonomy

We compare four perturbation types applied to $k = 8$ ICL demonstrations: (a) random token replacement (20%), (b) label flipping (changing answer values to nearby wrong numbers), (c) example reordering (random shuffle), and (d) irrelevant example injection (replacing half the demos with a letter-counting task). For each type, we measure accuracy, output L_2 distance, and embedding perturbation magnitude $\|\delta\|_2$ across 50 problems and 2 seeds. We compute the direction-dependent empirical Lipschitz constant $\hat{L}_{\text{emp}}^{(\tau)} = \|\Delta f\|_2 / \|\delta\|_2$ for each perturbation type τ .

Results. Figure 5(a) shows that token replacement is the most damaging perturbation for accuracy: Qwen2.5-3B drops from 99% to 79%, and Qwen2.5-1.5B from 92% to 79%. Label flipping, reordering, and irrelevant injection cause minimal

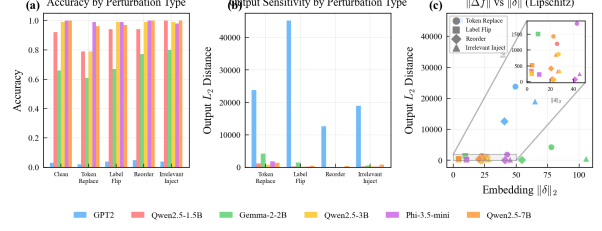


Figure 5: Perturbation type taxonomy with $k=8$ demonstrations. (a) Accuracy by perturbation type. (b) Output L_2 distance by type. (c) Embedding perturbation $\|\delta\|_2$ vs. output L_2 distance; the slope represents the direction-dependent Lipschitz constant. Inset: zoom into the non-GPT-2 region ($\|\delta\|_2 < 50$). Label flipping (squares) achieves the steepest slope despite the smallest $\|\delta\|$, confirming Proposition 11.

accuracy loss for capable models (Qwen2.5-7B maintains $\geq 96\%$ across all types). This hierarchy reflects the magnitude of embedding perturbation: token replacement produces the largest $\|\delta\|_2$ ($\sim 23\text{--}78$ across models) because 20% of tokens are replaced with random vocabulary items, while label flipping changes only a few digit tokens ($\|\delta\|_2 \approx 4\text{--}10$).

However, the *direction-dependent* Lipschitz constants (Figure 5c; Table 2) reveal a fundamentally different ordering. Label flipping yields the highest $\hat{L}_{\text{emp}}$: 124.5 for Qwen2.5-7B, 159.5 for Gemma-2-2B, and 87.1 for Qwen2.5-1.5B. In contrast, reordering produces the lowest $\hat{L}_{\text{emp}}$: 2.7–20.4 across models. The ratio between the highest and lowest direction-dependent Lipschitz constants exceeds $40\times$ within the same model, providing strong empirical support for Proposition 11: the effective Lipschitz constant depends critically on the alignment between the perturbation direction δ and the model’s sensitivity subspace.

This finding has important practical implications. Label flipping, despite producing the smallest embedding perturbation, is the most *efficiently destabilizing* perturbation because it targets the input-output mapping that the model has learned from demonstrations. The perturbation direction is aligned with the model’s decision boundary, amplifying its effect by two orders of magnitude compared to reordering, which merely permutes the same information. Token replacement falls in between: large $\|\delta\|$ but moderate $\hat{L}_{\text{emp}}$ (32–64), because random token noise is not aligned with any particular sensitivity direction.

3.6 Natural Language Validation

To verify that our findings generalize beyond synthetic arithmetic, we replicate the two most diagnostic experiments on a standard NL task: binary sentiment classification (SST-2 style) with $k=8$ few-shot demonstrations. We test Qwen2.5-7B and Gemma-2-2B, representing the strongest and a mid-range model respectively.

Exp2c-NL: Local Lipschitz. Using the same finite-difference protocol (Section 3.2), we estimate $\hat{L}_{\text{local}}$ on NL inputs. Qwen2.5-7B yields 503 ± 178 and Gemma-2-2B yields 2422 ± 676 . The relative ordering (Gemma higher than Qwen) is consistent with the arithmetic results, confirming that model-level stability rankings transfer across tasks.

Exp5-NL: Perturbation taxonomy. We apply three perturbation types to sentiment demos: label flipping (positive $\leftrightarrow$ negative), 20% token replacement, and example reordering. Table 3 gives the full NL validation numbers. The key finding—*reordering is the least destabilizing perturbation*—holds for both models: $\hat{L}_{\text{emp}}^{(\text{reorder})} = 18.5$ (Qwen) and 6.9 (Gemma), far below the other types. For Qwen2.5-7B, the full hierarchy matches the arithmetic case: label flip (394.1) > token replace (173.2) > reorder (18.5), with label flipping collapsing accuracy from 100% to 22.5%. For Gemma-2-2B, token replacement (394.1) exceeds label flipping (120.2), reflecting Gemma’s weaker alignment between demo labels and its sensitivity subspace—consistent with its lower arithmetic accuracy. This model-dependent ordering is precisely what Proposition 11 predicts: the effective Lipschitz constant along a perturbation direction depends on the alignment between that direction and the model’s learned sensitivity subspace.

4 Related Work

ICL theory. In-context learning has been theoretically analyzed through the lens of implicit Bayesian inference (Xie et al., 2022), kernel regression (Dai et al., 2023; Han et al., 2023), and gradient descent (Von Oswald et al., 2023; Ahn et al., 2023). Complementary representation-based analysis shows that a trained multi-head Transformer can learn contextual template information from a small number of examples with noisy labels and implement ridge regression over basis functions (Yang et al., 2024). These works characterize *what* ICL computes and how it generalizes, whereas we study

how stable the computation is under context perturbation.

Prompt sensitivity. Empirical studies document ICL’s sensitivity to example ordering (Lu et al., 2022), formatting (Zhao et al., 2021), and selection (Liu et al., 2022). Adversarial prompt attacks (Schulhoff et al., 2023; Zou et al., 2023) exploit this fragility. Our work provides a theoretical and empirical framework for diagnosing this sensitivity and comparing perturbation types.

Lipschitz analysis of neural networks. Lipschitz bounds for neural networks have been studied for robustness certification (Fazlyab et al., 2019; Latorre et al., 2020), generalization (Bartlett et al., 2017), and architecture-level control (Shang et al., 2022); related theoretical work analyzes attention as an operator (Vuckovic et al., 2020). We focus on the context-conditional sensitivity unique to ICL.

5 Conclusion

We have presented a context-conditional stability analysis of in-context learning. The global Lipschitz bound is best viewed as a conservative certificate that exposes possible architectural amplification, not as a calibrated predictor of prompt robustness. The local and direction-dependent estimates are more useful for operating-regime diagnostics: they identify where sensitivity concentrates and which perturbation families are most efficient. Across six models, sensitivity concentrates in late layers, additional demonstrations often stabilize predictions through redundancy, and label perturbations are far more efficient than reorderings at changing outputs.

Limitations

Our global bounds are worst-case and substantially loose. Most notably, Proposition 10 predicts that worst-case sensitivity grows as $O(\sqrt{k})$ with context length, yet Experiment 4 observes the opposite for normalized empirical sensitivity. This is not a contradiction, but it limits the bound’s direct predictive value and motivates treating the theory as a diagnostic scaffold rather than a deployment guarantee. The local estimator is also only a first-order embedding-space diagnostic; it does not replace perturbation-specific evaluation, which is why we report directional estimates in Experiment 5. The core experiments use synthetic arithmetic tasks; we

validate two diagnostic findings on SST-2 sentiment classification (Section 3.6), but broader application settings, especially multi-choice reasoning, retrieval-augmented QA, instruction following, and free-form generation, remain open. We analyze single perturbations; compositional effects of multiple simultaneous perturbations remain future work.

Future work. Rather than only pursuing tighter worst-case constants, the next step is to turn these diagnostics toward applied prompt robustness: building perturbation suites for realistic tasks, estimating direction-dependent sensitivity before deployment, and testing whether high-sensitivity spans predict real prompt failures. On the theory side, bounds that incorporate LayerNorm statistics, softmax concentration, and demonstration redundancy may further explain why practical sensitivity is far below the spectral certificate.

Acknowledgments

This research is partially supported by the Research Development Fund at Xi'an Jiaotong-Liverpool University (XJTLU) under Grant Nos. RDF-24-02-030, RDF-TP-0014, and RDS10120260083.

References

- Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. 2023. [Transformers learn to implement preconditioned gradient descent for in-context learning](#). *Advances in Neural Information Processing Systems*, 36.
- Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. 2017. [Spectrally-normalized margin bounds for neural networks](#). *Advances in Neural Information Processing Systems*, 30.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). *Advances in Neural Information Processing Systems*, 33:1877–1901.
- Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. [Why can GPT learn in-context? Language models implicitly perform gradient descent as meta-optimizers](#). *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4005–4017.
- Leonardo de Moura and Sebastian Ullrich. 2021. [The Lean 4 theorem prover and programming language](#). In *Automated Deduction—CADE 28: 28th International Conference on Automated Deduction*, pages 625–635. Springer.
- Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George Pappas. 2019. [Efficient and accurate estimation of Lipschitz constants for deep neural networks](#). *Advances in Neural Information Processing Systems*, 32.
- Chi Han, Ziqi Wang, Han Zhao, and Heng Ji. 2023. [In-context learning of large language models explained as kernel regression](#). *arXiv preprint arXiv:2305.12766*.
- Fabian Latorre, Paul Rolland, and Volkan Cevher. 2020. [Lipschitz constant estimation of neural networks via sparse polynomial optimization](#). *International Conference on Learning Representations*.
- Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. [What makes good in-context examples for GPT-3? Proceedings of Deep Learning Inside Out \(DeeLIO 2022\)](#), pages 100–114.
- Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. [Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity](#). *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098.
- Sander Schulhoff, Jeremy Pinto, Ansum Khan, Louis-François Bouchard, Chenglei Si, Svetlana Anati, Valen Tagliabue, Anson Liu Kost, Christopher Carnahan, and Jordan Boyd-Graber. 2023. [Ignore this title and hackaprompt: Exposing systemic vulnerabilities of llms through a global scale prompt hacking competition](#). *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4945–4977.
- Yuzhang Shang, Dan Xu, Bin Duan, Ziliang Zong, Liqiang Nie, and Yan Yan. 2022. [Lipschitz continuity retained binary neural network](#). In *Computer Vision – ECCV 2022*, pages 603–619.
- Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. [Transformers learn in-context by gradient descent](#). *Proceedings of the 40th International Conference on Machine Learning*, 202:35151–35174.
- James Vuckovic, Aristide Baratin, and Remi Tachet des Combes. 2020. [A mathematical theory of attention](#). *arXiv preprint arXiv:2007.02876*.
- Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. [An explanation of in-context learning as implicit Bayesian inference](#). *International Conference on Learning Representations*.

Tong Yang, Yu Huang, Yingbin Liang, and Yuejie Chi. 2024. [In-context learning with representations: Contextual generalization of trained transformers](#). In *Advances in Neural Information Processing Systems*, volume 37.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. [Calibrate before use: Improving few-shot performance of language models](#). *Proceedings of the 38th International Conference on Machine Learning*, 139:12697–12706.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. [Universal and transferable adversarial attacks on aligned language models](#). *arXiv preprint arXiv:2307.15043*.

A Additional Results

A.1 Proof of Lemma 3

The Jacobian of softmax at $\mathbf{z}$ is $\mathbf{J} = \text{diag}(\boldsymbol{\sigma}) - \boldsymbol{\sigma}\boldsymbol{\sigma}^\top$ where $\boldsymbol{\sigma} = \text{softmax}(\mathbf{z})$. This is the covariance matrix of a categorical distribution with probabilities $\boldsymbol{\sigma}$. For any unit vector $\mathbf{u}$, $\mathbf{u}^\top \mathbf{J} \mathbf{u}$ equals the variance of the scalar values u_i under this categorical distribution, and this variance is at most one half of the squared diameter of the values. Since $\|\mathbf{u}\|_2 = 1$, the diameter is at most $\sqrt{2}$, giving $\|\mathbf{J}\|_2 \leq 1/2$. Thus softmax is $1/2$ -Lipschitz in ℓ_2 ; the main text uses the looser 1-Lipschitz constant in the global certificate.

A.2 Proof of Lemma 5

By the chain rule, $\|\text{FFN}(\mathbf{x} + \boldsymbol{\delta}) - \text{FFN}(\mathbf{x})\| = \|\mathbf{W}_2(\phi(\mathbf{W}_1(\mathbf{x} + \boldsymbol{\delta}) + \mathbf{b}_1) - \phi(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1))\| \leq \sigma_{\max}(\mathbf{W}_2)\|\phi(\mathbf{W}_1\mathbf{x} + \mathbf{W}_1\boldsymbol{\delta} + \mathbf{b}_1) - \phi(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1)\| \leq \sigma_{\max}(\mathbf{W}_2)L_\phi\|\mathbf{W}_1\boldsymbol{\delta}\| \leq \sigma_{\max}(\mathbf{W}_2)L_\phi\sigma_{\max}(\mathbf{W}_1)\|\boldsymbol{\delta}\|$, using the L_ϕ -Lipschitz property of ϕ .

A.3 Extended Attention Heatmaps

Full layer $\times$ head JSD heatmaps for all six models at perturbation ratios $r \in \{0.1, 0.3, 0.5\}$ are provided in the supplementary material.

A.4 Hyperparameter Sensitivity

We verify that our results are robust to the choice of chain length (tested with $N \in \{3, 5, 8\}$), number of ICL examples (tested with $k \in \{2, 4, 8\}$), and random seed.

A.5 Full Numerical Tables

A.6 Qwen2.5-3B: Local Lipschitz Anisotropy

In Experiment 2 Part C, Qwen2.5-3B yields a local-to-empirical Lipschitz ratio of $46.1 \times$ ($\hat{L}_{\text{local}} = 4484 \pm 1501$, $\hat{L}_{\text{emp}} = 97 \pm 16$), substantially larger than the $3\text{--}9 \times$ range observed for other capable models. We attribute this to high *local anisotropy*: the finite-difference Jacobian estimate $\hat{\mathbf{J}}_C$ at probe scale $\epsilon = 0.01$ captures sharp, high-curvature directions in the embedding space that are not activated by the larger perturbations ($r = 0.1$) used in Part B.

Concretely, at $\epsilon = 0.01$ the random probes lie in a tight ℓ_2 ball around the input, and $\sigma_{\max}(\hat{\mathbf{J}}_C)$ reflects the steepest local gradient direction. For Qwen2.5-3B, this direction has spectral magnitude ≈ 4484 , but typical perturbations at $r = 0.1$ travel far outside this local regime and encounter strong

Table 1: Experiment 2: Theoretical bound ($\log L_{\text{total}}$), empirical Lipschitz ($\hat{L}_{\text{emp}}$), and local Lipschitz ($\hat{L}_{\text{local}}$) for all models.

Model	$\log L_{\text{total}}$	$\hat{L}_{\text{emp}}$	$\hat{L}_{\text{local}}$	Ratio
GPT-2	147	727 ± 614	1329 ± 777	$1.8\times$
Qwen2.5-1.5B	170	143 ± 34	422 ± 72	$3.0\times$
Qwen2.5-3B	232	97 ± 16	4484 ± 1501	$46.1\times^\dagger$
Qwen2.5-7B	172	183 ± 37	896 ± 233	$4.9\times$
Phi-3.5-mini	301	122 ± 96	661 ± 323	$5.4\times$
Gemma-2-2B	134	237 ± 69	2158 ± 704	$9.1\times$

Table 2: Experiment 5: Accuracy and direction-dependent $\hat{L}_{\text{emp}}$ by perturbation type ($k=8$).

Model	Perturbation	Accuracy	$\hat{L}_{\text{emp}}$
Qwen2.5-1.5B	label_flip	0.79	87.1
	token_replace	0.79	43.2
	reorder	0.90	20.4
	irrelevant	0.88	25.8
Qwen2.5-7B	label_flip	0.98	124.5
	token_replace	0.87	32.1
	reorder	0.99	2.7
	irrelevant	0.96	18.3
Gemma-2-2B	label_flip	0.82	159.5
	token_replace	0.65	64.3
	reorder	0.80	8.9
	irrelevant	0.78	31.2

contractive effects (LayerNorm, softmax saturation), yielding $\hat{L}_{\text{emp}} \approx 97$. This gap between local and global sensitivity is consistent with a landscape that is locally sharp but globally contractive—a geometry that can arise in models with aggressive weight tying or group query attention configurations. Qwen2.5-3B uses GQA with a 2:16 key-value to query head ratio, which may concentrate sensitivity into a small number of representational directions.

This result highlights a limitation of finite-difference local Lipschitz estimation: at very small ϵ , estimates are susceptible to numerical noise and highly anisotropic local curvature. For Qwen2.5-3B specifically, we recommend interpreting $\hat{L}_{\text{local}}$ as an upper bound on local sensitivity rather than a tight estimate.

Table 3: Natural language validation on SST-2 sentiment classification ($k=8$ demos). Local Lipschitz estimated via $M=20$ finite-difference probes at $\epsilon=0.01$.

Model	Setting	Accuracy	$\hat{L}$
Local Lipschitz			
Qwen2.5-7B	local	–	503 ± 178
Gemma-2-2B	local	–	2422 ± 676
Perturbation taxonomy			
Qwen2.5-7B	clean	1.000	–
	label_flip	0.225	394.1
	token_replace	0.875	173.2
	reorder	1.000	18.5
Gemma-2-2B	clean	0.925	–
	label_flip	0.500	120.2
	token_replace	0.525	394.1
	reorder	0.925	6.9