TCB: How Stable is the Next Token? A Geometric View of LLM Prediction Stability

Anonymous authors
Paper under double-blind review

Abstract

Large Language Models (LLMs) exhibit impressive capabilities yet suffer from sensitivity to slight input context variations, hampering reliability. Conventional metrics like accuracy and perplexity fail to assess local prediction robustness, as normalized output probabilities can obscure the underlying resilience of an LLM's internal state to perturbations. We introduce the **Token Constraint Bound** (δ_{TCB}), a novel metric that quantifies the maximum internal state perturbation an LLM can withstand before its dominant next-token prediction significantly changes. Intrinsically linked to output embedding space geometry, δ_{TCB} provides insights into the stability of the model's internal predictive commitment. Our experiments show δ_{TCB} correlates with effective prompt engineering and uncovers critical prediction instabilities missed by perplexity during in-context learning and text generation. δ_{TCB} offers a principled, complementary approach to analyze and potentially improve the contextual stability of LLM predictions.

1 Introduction

Large Language Models (LLMs), such as GPT-4 (OpenAI et al., 2023), LLaMA (Touvron et al., 2023; Dubey et al., 2024) and Gemini (Team et al., 2023), demonstrate remarkable capabilities, yet paradoxically exhibit striking sensitivity to contextual nuances. This brittleness manifests as substantial performance variations due to subtle modifications: accuracy can fluctuate by up to 76% from minor formatting changes (Sclar et al., 2023) or range from 54% to 93% based on example order (Zhao et al., 2021). Such variations stem from alterations in prompt phrasing (Razavi et al., 2025), example selection and ordering (Lu et al., 2021), or even basic formatting. Despite established scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022) fueling impressive in-context learning (Brown et al., 2020; Dong et al., 2022; Wei et al., 2023), evidence indicates that increased model scale does not inherently confer enhanced robustness; larger models may even exhibit new sensitivities (Lu et al., 2021; Wei et al., 2023). This underscores the urgent need for robust stability metrics in modern AI evaluation, particularly for reliable deployment in mission-critical applications demanding consistent performance (Weidinger et al., 2021; Herrera-Poyatos et al., 2025).

Appraising contextual influence with precision is imperative, yet existing evaluation frameworks prove inadequate. Task accuracy yields only an aggregate performance view, overlooking the stability of individual predictions amid contextual shifts. Perplexity (Jelinek et al., 1977), though standard for sequence likelihood (Liang et al., 2022; Holtzman et al., 2021), conflates probabilities, thereby obscuring local dynamics essential for robustness. Moreover, it often neglects internal state geometry and fails to ensure internal stability even for high-probability tokens (Cohen-Inger et al., 2025). Crucially, the softmax normalization applied to derive output probabilities can mask a prediction's underlying stability; high probability can arise from relative normalization, not necessarily from a robust internal state. This implies that a high token probability offers no guarantee that the originating internal state $\boldsymbol{h}$ is itself resilient to minor variations. Even as emerging metrics (Zhang et al., 2024; Tian et al., 2023; Geng et al., 2023) address confidence and calibrationchiefly by aligning probabilities with correctness likelihood (Tian et al., 2023)they do not directly gauge the robustness of a specific next-token prediction's dominant rank to perturbations in the internal representation $\boldsymbol{h}$. A well-calibrated, high-confidence prediction may therefore belie an unstable equilibrium within the internal state (Liu et al., 2025). This gap in assessing the immediate predictive mechanism's stability against internal perturbations is the direct impetus for our central research question:

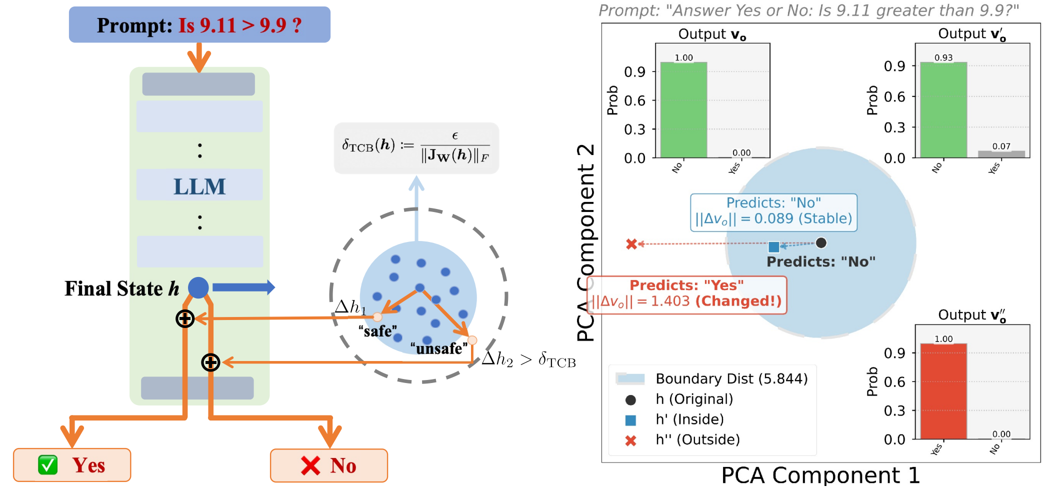


Figure 1: **The Token Constraint Bound (δ_{TCB}) mechanism.** δ_{TCB} quantifies the maximum perturbation a model's internal state can withstand before the next-token prediction changes. (a) **Left panel** illustrates how a hidden state perturbation $\Delta \boldsymbol{h}$ impacts the next token prediction. Small perturbations ($\Delta \boldsymbol{h}_1$, implicitly within δ_{TCB} radius) may preserve the output, while larger ones ($\Delta \boldsymbol{h}_2 > \delta_{\mathrm{TCB}}$) can flip it (from "No" to "Yes"). δ_{TCB} bounds the perturbation size for stable output. (b) **Right panel** shows that the original hidden state $\boldsymbol{h}$ and a perturbed state $\boldsymbol{h}'$ inside a stability region predict "No". Another perturbation $\boldsymbol{h}''$ outside the region flips the prediction to "Yes", demonstrating the practical consequence of exceeding the stability boundary.

> **$\mathbb{Q}$: How can we quantify the stability of an LLM's immediate prediction state, as induced by a specific prompt or context, against small internal variations?**

Addressing this question necessitates transcending aggregate performance metrics to develop measures specifically targeting the local robustness of prediction mechanisms. We must quantify how susceptible the next-token output distribution is to perturbations in the internal representation generated from the input contexta challenge at the intersection of representation stability and prediction reliability.

Our approach. We propose the **Token Constraint Bound** (δ_{TCB}), a measure of this critical local stability. δ_{TCB} quantifies a "safety margin" around the internal state $\boldsymbol{h}$ resulting from context processing: a larger δ_{TCB} means the models next-token prediction (particularly its top choice) withstands greater internal perturbations $\Delta \boldsymbol{h}$ without significant change. It gauges the model's commitment to its current output ranking, given $\boldsymbol{h}$. As explicated in Section 2 and depicted in Figure 1, δ_{TCB} offers a direct measure of the output layer's robustness to hidden state variations. Therefore, a high δ_{TCB} signals a *stably* confident prediction state engendered by effective context.

We hypothesize that effective context, such as well-crafted prompts or informative ICL examples, not only guides models to correct answers but also induces a more *stable* internal state $\boldsymbol{h}$, as reflected by higher δ_{TCB} values. This stability, signifying robust internal commitment to a predictive path, yields more reliable predictions. Consequently, δ_{TCB} offers a quantitative measure for context effectiveness beyond accuracy, serving as a proxy for the robustness of context-derived decision-making and complementing uncertainty metrics focused on "knowledge strength" (Ma et al., 2025).

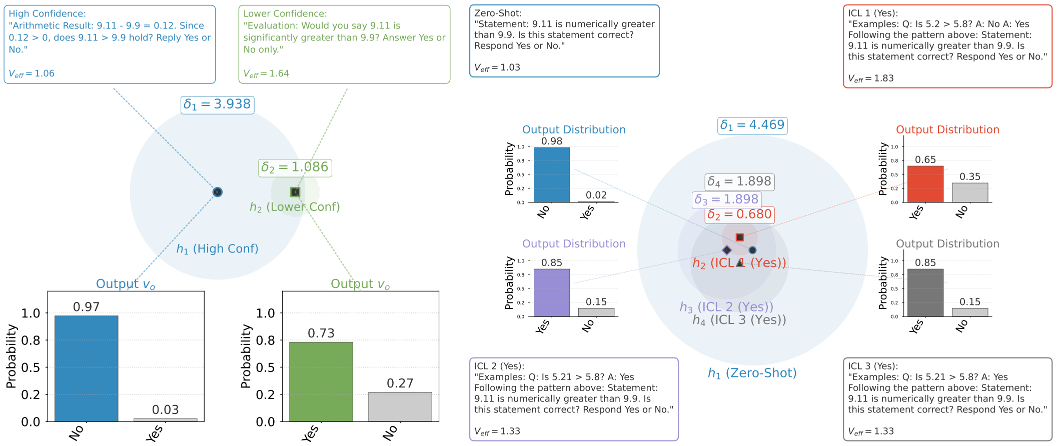
Our experiments corroborate this. We show δ_{TCB} distinguishes prompt quality and exhibits distinct behaviors across confidence regimes, correlating with distributional flatness in uncertain cases and logit margins in high-confidence scenarios. Results confirm δ_{TCB}'s sensitivity to output embedding geometry, its link to semantic content, and its ability to flag incipient instability during text generationdynamics perplexity overlooks.

Our contributions are threefold:

(a) We introduce and theoretically ground the Token Constraint Bound (δ_{TCB}), a novel metric that measures the local robustness of an LLM's next-token prediction to internal state perturbations, and detail its practical computation (Section 2).

(b) We derive an expression that intrinsically links δ_{TCB} to the geometric dispersion of output embeddings, thus identifying geometric underpinnings of prediction stability (Section 3).

(c) Through empirical evaluation, we demonstrate δ_{TCB}'s capacity to assess prompt effectiveness and showcase its application in refining both prompt engineering and ICL (Section 4).



High Confidence:
"Arithmetic Result: 9.11 - 9.9 = 0.12. Since 0.12 > 0, does 9.11 > 9.9 hold? Reply Yes or No."
$V_{eff} = 1.06$

Lower Confidence:
"Evaluation: Would you say 9.11 is significantly greater than 9.9? Answer Yes or No only."
$V_{eff} = 1.64$

Zero-Shot:
"Statement: 9.11 is numerically greater than 9.9. Is this statement correct? Respond Yes or No."
$V_{eff} = 1.03$

ICL 1 (Yes):
"Examples: Q: Is 5.2 > 5.8? A: No A: Yes Following the pattern above: Statement: 9.11 is numerically greater than 9.9. Is this statement correct? Respond Yes or No."
$V_{eff} = 1.83$

ICL 2 (Yes):
"Examples: Q: Is 5.21 > 5.8? A: Yes Following the pattern above: Statement: 9.11 is numerically greater than 9.9. Is this statement correct? Respond Yes or No."
$V_{eff} = 1.33$

ICL 3 (Yes):
"Examples: Q: Is 5.21 > 5.8? A: Yes Following the pattern above: Statement: 9.11 is numerically greater than 9.9. Is this statement correct? Respond Yes or No."
$V_{eff} = 1.33$

(a) Effect of prompt confidence on δ_{TCB}.

(b) Effect of In-Context Learning on δ_{TCB}.

Figure 2: δ_{TCB} **reflects context-induced prediction stability.** (a) Illustrates how prompts inducing higher prediction confidence (lower $\mathcal{V}_{\text{eff}}$, state $\boldsymbol{h}_1$) lead to a significantly larger δ_{TCB} compared to prompts yielding lower confidence (higher $\mathcal{V}_{\text{eff}}$, state $\boldsymbol{h}_2$). (b) Shows how In-Context Learning examples modify the hidden state and consequently the prediction and its stability. Adding examples can initially decrease stability while flipping the prediction, but consistent examples can increase stability for the target output.

2 Preliminaries: Understanding LLM Predictions and Stability

This section establishes the foundational concepts for analyzing local output stability in LLMs. We begin by outlining the LLM output mechanism, then explore what it means for a prediction to be stable against internal variations. This leads to the introduction of the Jacobian matrix as a tool for quantifying sensitivity, and finally culminates in the formal definition of our δ_{TCB}.

2.1 Language Model Output and Distribution Concentration

Consider an LLM whose final layer computes a hidden state $\boldsymbol{h} \in \mathbb{R}^d$. This state is linearly transformed by an output weight matrix $\mathbf{W} \in \mathbb{R}^{\mathcal{V} \times d}$ to produce logits $\boldsymbol{z} = \mathbf{W}\boldsymbol{h}$, where $\mathcal{V}$ is the vocabulary size. Each row $\boldsymbol{w}_i^\top$ of $\mathbf{W}$ corresponds to the output embedding for token i. The probability distribution over the next token, $\boldsymbol{o} \in \mathbb{R}^{\mathcal{V}}$, is obtained via the softmax function:

$$\boldsymbol{o} = \operatorname{softmax}(\boldsymbol{z}), \quad \text{where } o_i = \frac{\exp(z_i)}{\sum_{j=1}^{\mathcal{V}} \exp(z_j)}. \tag{1}$$

This distribution satisfies $\sum_{i=1}^{\mathcal{V}} o_i = 1$ and $o_i \geq 0$. The model's prediction is typically the token i^* maximizing o_i. A useful measure of the concentration of this distribution is the *effective vocabulary size* $\mathcal{V}_{\text{eff}}$:

$$\mathcal{V}_{\text{eff}}(\boldsymbol{o}) := \frac{1}{\sum_{i=1}^{\mathcal{V}} o_i^2} = \frac{1}{\|\boldsymbol{o}\|_2^2}. \tag{2}$$

$\mathcal{V}_{\text{eff}}$ ranges from 1 to $\mathcal{V}$, inversely relating to the L_2 norm squared of the probability vector. Our analysis hinges on understanding how the characteristics of this output vector $\boldsymbol{o}$, including its concentration, relate to its *stability* when the context-derived hidden state $\boldsymbol{h}$ undergoes small changes.

2.2 Defining Stability: What Does It Mean for a Prediction to Be Stable?

Our core objective is to understand the robustness of an LLM's next-token prediction. Specifically, we want to know: if the LLM's internal summary of the context (represented by the final hidden state $\boldsymbol{h}$) changes slightly, how much does its next-token probability distribution $\boldsymbol{o}$ change? Let $\Delta \boldsymbol{h} \in \mathbb{R}^d$ represent a small internal "wobble" or perturbation to the hidden state $\boldsymbol{h}$. Such a perturbation to the *model's internal representation of the context* could arise from minor input variations, noise in the computation, or other subtle disturbances. Let $\boldsymbol{o}' = \operatorname{softmax}(\mathbf{W}(\boldsymbol{h} + \Delta \boldsymbol{h}))$ be the perturbed output distribution. The resulting change in the prediction is $\Delta \boldsymbol{o} = \boldsymbol{o}' - \boldsymbol{o}$.

The central question motivating our work, reiterated from the Introduction $\mathbb{Q}$, is how to quantify a "safety margin" for $\boldsymbol{h}$: how large can the perturbation $\Delta \boldsymbol{h}$ be before the change in the output $\Delta \boldsymbol{o}$ becomes unacceptably large? Answering this requires a way to relate the magnitude of the internal perturbation $\Delta\boldsymbol{h}$ to the magnitude of the resulting output change $\Delta\boldsymbol{o}$. This safety margin offers insights distinct from interpreting output probabilities $\boldsymbol{o}$ as direct measures of absolute confidence; instead, δ_{TCB} focuses on the *local integrity and resilience of the current predictive mechanism* itself.

2.3 Quantifying the Impact of Perturbations: The Role of the Jacobian

To precisely relate changes in $\boldsymbol{h}$ to changes in $\boldsymbol{o}$, we utilize the concept of the Jacobian matrix. To first order, for small $\Delta\boldsymbol{h}$, the change in the output distribution $\Delta\boldsymbol{o}$ can be approximated linearly:

$$\Delta\boldsymbol{o} \approx \mathbf{J}_{\mathbf{W}}(\boldsymbol{h})\Delta\boldsymbol{h}, \tag{3}$$

where $\mathbf{J}_{\mathbf{W}}(\boldsymbol{h}) \in \mathbb{R}^{\mathcal{V}\times d}$ is the Jacobian matrix of the output probabilities $\boldsymbol{o}$ with respect to the hidden state $\boldsymbol{h}$. It is given by:

$$\mathbf{J}_{\mathbf{W}}(\boldsymbol{h}) = \frac{\partial\boldsymbol{o}}{\partial\boldsymbol{h}} = \underbrace{\frac{\partial\boldsymbol{o}}{\partial\boldsymbol{z}}}_{\mathrm{diag}(\boldsymbol{o})-\boldsymbol{o}\boldsymbol{o}^{\top}} \underbrace{\frac{\partial\boldsymbol{z}}{\partial\boldsymbol{h}}}_{\mathbf{W}} = \left(\mathrm{diag}(\boldsymbol{o}) - \boldsymbol{o}\boldsymbol{o}^{\top}\right)\mathbf{W}. \tag{4}$$

The Jacobian $\mathbf{J}_{\mathbf{W}}(\boldsymbol{h})$ essentially captures the sensitivity of each output probability o_i to infinitesimal changes in each dimension of the hidden state h_k. Its entries are $\frac{\partial o_i}{\partial h_k}$. Note that the Jacobian depends on both the current output distribution $\boldsymbol{o}$ and the output weight matrix $\mathbf{W}$. To relate the overall magnitude of the state perturbation $\|\Delta\boldsymbol{h}\|_2$ to the overall magnitude of the output change $\|\Delta\boldsymbol{o}\|_2$, we use matrix norms. A standard inequality bounds the output change using the Jacobian's Frobenius norm:

$$\|\Delta\boldsymbol{o}\|_2 \leq \|\mathbf{J}_{\mathbf{W}}(\boldsymbol{h})\|_F\|\Delta\boldsymbol{h}\|_2. \tag{5}$$

The Frobenius norm $\|\mathbf{J}_{\mathbf{W}}(\boldsymbol{h})\|_F = \left(\sum_{i=1}^{\mathcal{V}}\sum_{k=1}^{d}(\frac{\partial o_i}{\partial h_k})^2\right)^{1/2}$ provides a comprehensive, aggregate measure of the sensitivity of all output probabilities to all hidden state dimensions. A larger Frobenius norm indicates that, the output probabilities are more sensitive to changes in the hidden state.

2.4 The Token Constraint Bound (δ_{TCB}): Our Measure of Stability

We are interested in finding the maximum allowable perturbation radius $\|\Delta\boldsymbol{h}\|_2$ such that the resulting change in the output distribution, as measured by its L_2 norm $\|\Delta\boldsymbol{o}\|_2$, remains below a predefined small tolerance $\epsilon > 0$. That is, we impose the condition $\|\Delta\boldsymbol{o}\|_2 \leq \epsilon$. Using the bound from Eq. (5), we require $\|\mathbf{J}_{\mathbf{W}}(\boldsymbol{h})\|_F\|\Delta\boldsymbol{h}\|_2 \leq \epsilon$. Rearranging for $\|\Delta\boldsymbol{h}\|_2$ gives us:

$$\|\Delta\boldsymbol{h}\|_2 \leq \frac{\epsilon}{\|\mathbf{J}_{\mathbf{W}}(\boldsymbol{h})\|_F}. \tag{6}$$

This naturally motivates our core metric for local output stability, which we define as this upper bound on the perturbation norm:

Definition 1 (Token Constraint Bound δ_{TCB}) . *Given the output weight matrix $\mathbf{W}$, hidden state $\boldsymbol{h}$, resulting output distribution $\boldsymbol{o} = \mathit{softmax}(\mathbf{W}\boldsymbol{h})$, and a tolerance $\epsilon > 0$ for the maximum L_2 change allowed in $\boldsymbol{o}$, the* Token Constraint Bound δ_{TCB} *at state $\boldsymbol{h}$ is defined as:*

$$\delta_{\mathrm{TCB}}(\boldsymbol{h}) := \frac{\epsilon}{\|\mathbf{J}_{\mathbf{W}}(\boldsymbol{h})\|_F}. \tag{7}$$

Here, $\mathbf{J}_{\mathbf{W}}(\boldsymbol{h})$ is the Jacobian given by Eq. (4) *and $\|\cdot\|_F$ denotes the Frobenius norm.*

The parameter ϵ is a dimensionless scale factor chosen by the user, representing the desired tolerance for output distribution change. Consequently, $\delta_{\mathrm{TCB}}(\boldsymbol{h})$ quantifies the L_2-norm radius of the largest hyper-sphere of perturbations $\Delta\boldsymbol{h}$ around the current hidden state $\boldsymbol{h}$ that, to a first-order approximation, guarantees the change in the output probability vector $\boldsymbol{o}$ remains within ϵ. A larger $\delta_{\mathrm{TCB}}(\boldsymbol{h})$ signifies that the model's *current prediction state* $\boldsymbol{o}$, as induced by the context leading to $\boldsymbol{h}$, is intrinsically more robust to small internal variations ("wobbles") in this hidden state. Conversely, a smaller $\delta_{\mathrm{TCB}}(\boldsymbol{h})$ indicates that the prediction mechanism is more sensitive to such internal perturbations at this specific point. The crucial term $\|\mathbf{J}_{\mathbf{W}}(\boldsymbol{h})\|_F^2$ in the denominator has a fundamental connection to the geometry of the output embeddings, which we explore in detail in Section 3.

3 δ_{TCB} via Output Embedding Geometry

Def. 1 introduced the Token Constraint Bound (δ_{TCB}) as a measure of local output stability. To unlock its full diagnostic power and understand its nuanced behavior, we now dissect its core component: the Frobenius norm of the output Jacobian, $\|\mathbf{J}_{\mathbf{W}}(\boldsymbol{h})\|_F$. This section reveals that $\|\mathbf{J}_{\mathbf{W}}(\boldsymbol{h})\|_F$, and consequently δ_{TCB}, is deeply intertwined with the *geometric arrangement* of the model's output token embeddings $\boldsymbol{w}_i$ relative to the current prediction probabilities $\boldsymbol{o}$. Intuitively, a prediction is expected to be more stable if the leading token's embedding is well-isolated from competitors, or if the model is highly certain (peaked $\boldsymbol{o}$).

3.1 The Geometry of Output: Token Embeddings and Their Mean

Recall that the output weight matrix $\mathbf{W} \in \mathbb{R}^{\mathcal{V}\times d}$ contains the output embedding vector $\boldsymbol{w}_i^\top$ for each token i as its rows. A key concept in understanding the geometric influence on stability is the probability-weighted mean embedding vector:

$$\boldsymbol{\mu}_{\boldsymbol{w}}(\boldsymbol{h}) := \sum_{j=1}^{\mathcal{V}} o_j \boldsymbol{w}_j = \mathbf{W}^\top \boldsymbol{o}. \tag{8}$$

Here, $\boldsymbol{\mu}_{\boldsymbol{w}}(\boldsymbol{h}) \in \mathbb{R}^d$ represents the current probability-weighted locus within the embedding space. This mean vector, $\boldsymbol{\mu}_{\boldsymbol{w}}(\boldsymbol{h})$, reflects the current probability-weighted locus within the embedding space, effectively representing the "center of mass" or the resultant directional influence of the entire output distribution on the embedding geometry.

3.2 Deriving the Jacobian Norm: Connecting Sensitivity to Embedding Spread

We now derive an exact analytical expression for the squared Frobenius norm of the output Jacobian, $\|\mathbf{J}_{\mathbf{W}}(\boldsymbol{h})\|_F^2$, which is the crucial term determining δ_{TCB}. Let $\{\boldsymbol{w}_i\}_{i=1}^{\mathcal{V}}$ be the output embedding vectors (rows of $\mathbf{W}$) and $\boldsymbol{\mu}_{\boldsymbol{w}}(\boldsymbol{h})$ be the probability-weighted mean embedding as defined in Eq. (8).

Proposition 1 (Exact Squared Jacobian Norm Appendix I) . *For a given output weight matrix* $\mathbf{W}$ *and hidden state* $\boldsymbol{h}$*, let* $\boldsymbol{o} = \mathit{softmax}(\mathbf{W}\boldsymbol{h})$ *be the output probability vector. The squared Frobenius norm of the output Jacobian* $\mathbf{J}_{\mathbf{W}}(\boldsymbol{h}) = (\mathrm{diag}(\boldsymbol{o}) - \boldsymbol{o}\boldsymbol{o}^\top)\mathbf{W}$ *is exactly:*

$$\|\mathbf{J}_{\mathbf{W}}(\boldsymbol{h})\|_F^2 = \sum_{i=1}^{\mathcal{V}} o_i^2 \, \|\boldsymbol{w}_i - \boldsymbol{\mu}_{\boldsymbol{w}}(\boldsymbol{h})\|_2^2 \, . \tag{9}$$

This sum represents the squared Euclidean distances between each embedding $\boldsymbol{w}_i$ *and the mean embedding* $\boldsymbol{\mu}_{\boldsymbol{w}}(\boldsymbol{h})$*, weighted by the corresponding squared probability* o_i^2*.*

Interpretation of the Formula. Eq. (9) is pivotal. It states that the overall sensitivity of the output distribution to hidden state perturbations (as captured by $\|\mathbf{J}_{\mathbf{W}}(\boldsymbol{h})\|_F^2$) is determined by how "spread out" the token embeddings $\boldsymbol{w}_i$ are from their probability-weighted mean $\boldsymbol{\mu}_{\boldsymbol{w}}(\boldsymbol{h})$, with each squared distance $\|\boldsymbol{w}_i - \boldsymbol{\mu}_{\boldsymbol{w}}(\boldsymbol{h})\|_2^2$ being amplified or diminished by the square of its token's probability o_i^2. The o_i^2 weighting is crucial:

- Embeddings for tokens with very low probability o_i (and thus low current ""evidence” or ""belief” from the model) contribute minimally to the sum, even if geometrically distant from $\boldsymbol{\mu}_{\boldsymbol{w}}$. The model effectively de-weights their geometric influence on stability at this state.
- Embeddings for high-probability tokens (carrying significant ""evidence”) contribute substantially, particularly if they are far from $\boldsymbol{\mu}_{\boldsymbol{w}}$. Their geometric influence on the Jacobian norm is quadratically emphasized by o_i^2.

This o_i^2 weighting distinguishes Eq. (9) from measures like the trace of the standard probability-weighted covariance matrix of embeddings. This distinction arises directly from the definition of the softmax Jacobian and is fundamental for correctly interpreting δ_{TCB} (see Appendix G).

3.3 The Full Form of δ_{TCB} and Its Geometric Meaning

Substituting the exact squared Jacobian norm from Prop. 1 Eq. (9) into the definition of δ_{TCB} Eq. (7) yields its complete form:

$$\delta_{\mathrm{TCB}}(\boldsymbol{h}) = \frac{\epsilon}{\sqrt{\sum_{i=1}^{\mathcal{V}} o_i^2 \, \|\boldsymbol{w}_i - \boldsymbol{\mu}_{\boldsymbol{w}}(\boldsymbol{h})\|_2^2}}. \tag{10}$$

Table 1: **Pearson Correlations validating regime-dependent stability.** Shows strong positive Corr($\delta_{\text{TCB}}, \mathcal{V}_{\text{eff}}$) (**r=0.95**) in diverse prompts (DPD, $N = 309$, broad regime), indicating stability driven by flatness. In contrast, strong positive Corr($\delta_{\text{TCB}}, z_k - z_{j^*}$) (**r=0.62**) emerges in high-confidence cases (Low-$\mathcal{V}_{\text{eff}}$ Targeted, LVD, $N = 360$), where Corr($\delta_{\text{TCB}}, \mathcal{V}_{\text{eff}}$) is negligible ($r = 0.08$), confirming stability relies on top-token separation when confidence is high. Metrics are the Token Constraint Bound (δ_{TCB}), effective vocabulary size ($\mathcal{V}_{\text{eff}}$), and the logit margin between the top two tokens ($z_k - z_{j^*}$).

Dataset (N Samples)	Corr($\delta_{\text{TCB}}, \mathcal{V}_{\text{eff}}$)	Corr($\delta_{\text{TCB}}, z_k - z_{j^*}$)	Corr($z_k - z_{j^*}, \mathcal{V}_{\text{eff}}$)
Diverse Prompts (DPD, $N = 309$)	**0.95** (Strong +)	-0.40 (Moderate -)	-0.41 (Moderate -)
Low-$\mathcal{V}_{\text{eff}}$ Targeted (LVD, $N = 360$)	0.08 (Near Zero)	**0.62** (Strong +)	-0.60 (Strong -)

This equation provides a clear geometric interpretation: δ_{TCB} is inversely proportional to the square root of the o_i^2-weighted sum of squared Euclidean distances between each token embedding $\boldsymbol{w}_i$ and the probability-weighted mean embedding $\boldsymbol{\mu}_{\boldsymbol{w}}(\boldsymbol{h})$. A larger δ_{TCB} indicates higher local robustness of the prediction generated from $\boldsymbol{h}$. This geometric dispersion, weighted by o_i^2, directly dictates the "safety radius" around $\boldsymbol{h}$, within which the output distribution $\boldsymbol{o}$ changes by at most ϵ. Understanding this relationship is key to interpreting how context shapes δ_{TCB} and, by extension, prediction stability, as conceptually illustrated in Figure 1 and Figure 3.

3.4 Interpreting Stability Across Prediction Regimes

The exact expression for $\|\mathbf{J_W}(\boldsymbol{h})\|_F^2$ in Eq. (9) and the visualization in Figure 3 elucidate how δ_{TCB} behaves under different prediction certainties.

High-Confidence Regime (Low $\mathcal{V}_{\text{eff}}$, Peaked $\boldsymbol{o}$). When $\boldsymbol{o}$ is highly peaked on token k ($o_k \to 1, \mathcal{V}_{\text{eff}} \to 1$), $\boldsymbol{\mu}_{\boldsymbol{w}}(\boldsymbol{h}) \to \boldsymbol{w}_k$, causing $\|\mathbf{J_W}(\boldsymbol{h})\|_F^2 \to 0$ and $\delta_{\text{TCB}} \to \infty$ (Figure 3 a). Here, extreme certainty implies extreme stability. The sum approximates to $\sum_{j \neq k} o_j^2 \|\boldsymbol{w}_j - \boldsymbol{w}_k\|_2^2$ (Appendix J). Crucially, competitor probabilities o_j (and thus o_j^2) are super-exponentially sensitive to the logit margin between the top-two candidates, $z_k - z_{j^*} = z_{\text{top1}} - z_{\text{top2}}$. Larger $z_k - z_{j^*}$ values drastically reduce $\|\mathbf{J_W}(\boldsymbol{h})\|_F^2$, boosting δ_{TCB}. This underpins the empirical positive correlation between δ_{TCB} and $z_k - z_{j^*}$ when $\mathcal{V}_{\text{eff}}$ is low (Table 1, Low-$\mathcal{V}_{\text{eff}}$). Distances $\|\boldsymbol{w}_j - \boldsymbol{w}_k\|_2^2$ further modulate this: more distant competitors require even smaller o_j^2 for the same stability.

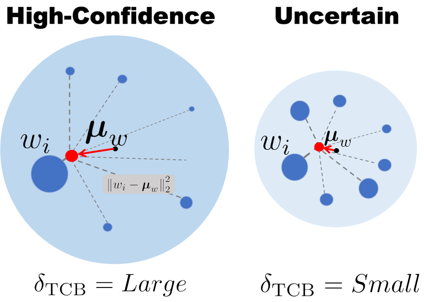


Figure 3: **Output Distribution Determines Geometric Stability.** The Token Constraint Bound (δ_{TCB}) is a function of the geometric arrangement of embeddings. (a) **High Confidence:** Peaked distribution concentrates $\boldsymbol{\mu}_{\boldsymbol{w}}(\boldsymbol{h})$ near the dominant embedding $\boldsymbol{w}_k$, minimizing the sum and maximizing δ_{TCB}. (b) **Uncertainty:** Flatter distribution spreads $\boldsymbol{\mu}_{\boldsymbol{w}}(\boldsymbol{h})$ among active embeddings, increasing the sum and reducing δ_{TCB}. $\boldsymbol{\mu}_{\boldsymbol{w}}(\boldsymbol{h})$, token embeddings $\boldsymbol{w}_i$.

Uncertain Regime (Higher $\mathcal{V}_{\text{eff}}$, Flatter $\boldsymbol{o}$). When probability is spread over multiple tokens (larger $\mathcal{V}_{\text{eff}}$, Figure 3 b), many o_i are non-negligible. If these probable tokens' embeddings $\boldsymbol{w}_i$ are distant from $\boldsymbol{\mu}_{\boldsymbol{w}}(\boldsymbol{h})$, their $\|\boldsymbol{w}_i - \boldsymbol{\mu}_{\boldsymbol{w}}(\boldsymbol{h})\|_2^2$ terms increase $\|\mathbf{J_W}(\boldsymbol{h})\|_F^2$, reducing δ_{TCB}. Crucially, however, a high $\mathcal{V}_{\text{eff}}$ does not *guarantee* low δ_{TCB}: if high-probability embeddings $\{\boldsymbol{w}_i\}$ are geometrically *clustered* (and thus all near $\boldsymbol{\mu}_{\boldsymbol{w}}(\boldsymbol{h})$), the $\|\boldsymbol{w}_i - \boldsymbol{\mu}_{\boldsymbol{w}}(\boldsymbol{h})\|_2^2$ terms could be small despite significant o_i^2 values, potentially resulting in a larger δ_{TCB}. This highlights geometry's primacy, validated by experiments where clustering embeddings (fixed $\boldsymbol{o}$) increases δ_{TCB}. In this uncertain regime, under simplifying assumptions (Appendix F, Appendix J), approximations can suggest $\|\mathbf{J_W}(\boldsymbol{h})\|_F^2 \propto 1/\mathcal{V}_{\text{eff}}$. This implies $\delta_{\text{TCB}} \propto \sqrt{\mathcal{V}_{\text{eff}}}$, aligning with empirical correlations over diverse prompts (Table 1, Diverse Prompts), where overall distribution shape often dominates individual logit margins.

4 Experiments

This section empirically substantiates the theoretical framework for δ_{TCB}, focusing on its connection to output embedding geometry and its utility in LLM analysis. We address:

- δ_{TCB}'s intrinsic properties, including sensitivity to output embedding geometry and correlations with standard metrics ($\mathcal{V}_{\text{eff}}, z_k - z_{j^*}$) across confidence regimes (Section 4.2).
- The role in diagnosing accuracy-stability conflicts and robust prompt engineering (Section 4.3).
- How δ_{TCB} complements Perplexity (PPL) by assessing local prediction robustness (Section D.3).

4.1 Experimental Setup

- ***Models.*** Primary experiments utilize the LLaMA-3.1–8B model (Touvron et al., 2023). All computations were performed on NVIDIA RTX 4090 GPUs.
- ***Datasets and Rationale.*** **MMLU (Hendrycks et al., 2020):** Employed for broad validation and initial characterization due to its diverse subject matter. We typically sampled $N_{\text{init\_pool}} = 100$ questions from "test" splits of 3-5 reasoning-heavy subjects (e.g., `formal_logic`, `philosophy`) to assess general trends and robustness under varied content. **GSM8K (Cobbe et al., 2021):** Utilized for detailed intervention analysis and prompt optimization case studies. Its multi-step reasoning nature provides a fertile ground for examining how nuanced prompt changes affect both accuracy and internal stability. An initial pool of $N_{\text{init\_pool}} = 100$ questions from the "test" set was used for broader studies, with specific problems selected for deep dives. **DPD and LVD Datasets:** We synthesized two datasets for correlation analysis. The **Diverse Prompts Dataset (DPD)** contains prompts from a range of tasks, designed to elicit varied model confidence levels. The **Low-$\mathcal{V}_{\text{eff}}$ Targeted Dataset (LVD)** was created by modifying DPD prompts to generate high-confidence predictions.
- ***Prompting Strategies.*** *Zero-Shot:* Minimalist prompts for baseline correlation studies and initial conflict identification. *Few-Shot / Interventions:* A range of k-shot prompts ($k = 5$), ICL variations (e.g., algebraic vs. arithmetic focus, hyper-specific examples), and instructional prefixes were used in diagnostic analyses and optimization experiments.
- ***Core Metrics.*** At critical prediction points (e.g., just before generating the answer token for MMLU multiple-choice; before the first token of the final numerical answer in GSM8K): **Token Constraint Bound ($\delta_{\text{TCB}}(\boldsymbol{h})$):** Computed via Eq. (10). For all experiments, we set the tolerance parameter $\epsilon = 1.0$, which normalizes the metric. Since our analysis focuses on *relative changes* in stability, the specific value of ϵ is less critical than its consistency. Higher δ_{TCB} indicates greater internal state robustness. **Effective Vocabulary Size ($\mathcal{V}_{\text{eff}}(\boldsymbol{o})$):** From Eq. (2). Lower $\mathcal{V}_{\text{eff}}$ indicates a more peaked, confident distribution. **Logit Margin ($z_k - z_{j^*}$):** Defined as $z_{\text{top1}} - z_{\text{top2}}$. Larger positive values suggest stronger discrimination for the top choice. **Task-Specific Accuracy (Acc):** Binary score based on ground truth. **Perplexity (PPL):** For local analysis in relation to δ_{TCB}, we often refer to the negative log probability of the predicted token ($-\log o_{\text{predicted}}$). Further details on experimental configurations are in Appendix C .

4.2 Intrinsic Properties and Validation of δ_{TCB}

Validating Sensitivity to Output Embedding Geometry Objective: To empirically confirm δ_{TCB}'s direct dependence on output embedding geometry ($\mathbf{W}$), independent of the probability distribution ($\boldsymbol{o}$). **Method Summary:** We synthetically manipulated $\mathbf{W}$ (clustering/dispersing competitor embeddings) while holding $\boldsymbol{h}$ and $\boldsymbol{o}$ (thus local PPL) constant for diverse MMLU prompts. δ_{TCB} was recalculated. **Results:** Table 2 shows the hypothesis $\delta_{\text{TCB}}(\mathbf{W}_{\text{cluster}}) > \delta_{\text{TCB}}(\mathbf{W}_{\text{orig}}) > \delta_{\text{TCB}}(\mathbf{W}_{\text{disperse}})$ held for **90% of prompts overall**, directly substantiating the geometric term in Eq. (10). This highlights that δ_{TCB} captures a dimension of stability tied to the embedding space that probability-only metrics would miss.

Table 2: **Simulation confirms δ_{TCB}'s geometric sensitivity.** Percentage of prompts validating $\delta_{\text{cluster}} > \delta_{\text{orig}} > \delta_{\text{disperse}}$ when manipulating $\mathbf{W}$ ($K = 10$ competitors) while fixing $\boldsymbol{o}$. The **effect robustly held (90% overall)**, confirming geometric influence distinct from probability shape.

Prompt Category	Hypothesis Held
Low $\mathcal{V}_{\text{eff}}$ (< 20)	95%
Medium $\mathcal{V}_{\text{eff}}$ (20-100)	92%
High $\mathcal{V}_{\text{eff}}$ (> 100)	80%
Overall	**90%**

Correlations Across Different Confidence Regimes **Objective:** To validate predicted shifts in δ_{TCB}'s correlations with $\mathcal{V}_{\text{eff}}$ and $z_k - z_{j^*}$ based on prediction confidence. **Method Summary:** Two MMLU zero-shot datasets: Diverse Prompts (DPD, $N = 309$) and Low-$\mathcal{V}_{\text{eff}}$ Targeted (LVD, $N = 360$). **Results:** Table 1 confirms the theorized regime dependence. In the DPD (broad regime), **Corr($\delta_{\text{TCB}}, \mathcal{V}_{\text{eff}}$) is** 0.95, indicating stability is largely driven by overall distribution flatness. In stark contrast, for the LVD (high-confidence), Corr($\delta_{\text{TCB}}, \mathcal{V}_{\text{eff}}$) drops to a negligible 0.08, while **Corr($\delta_{\text{TCB}}, z_k - z_{j^*}$) becomes a strong** 0.62. This shift empirically validates that when the model is confident, δ_{TCB} reflects the separation of the top token from its competitors rather than just the general peakedness of the distribution.

4.3 δ_{TCB} as a Diagnostic Tool for Prompt Engineering

δ_{TCB} uniquely identifies stability issues that accuracy or conventional confidence metrics ($\mathcal{V}_{\text{eff}}, z_k - z_{j^*}$) may miss. Common *accuracy-stability conflict scenarios* include: (1) **Accurate but Unstable**:

Table 3: Combined Impact of δ_{TCB}-Enhancement on Mean Metrics for Unperturbed and Perturbed Prompts (MMLU & GSM8K). Metrics are Accuracy (Acc), Token Constraint Bound (δ_{TCB}), Effective Vocabulary Size ($\mathcal{V}_{\text{eff}}$), and Logit Margin ($z_k - z_{j^*}$). Perturbed metrics include Accuracy Variance ($\text{AccVar}_{\text{pert}}$), Performance Drop Rate (PDR), and Worst-Case Accuracy ($\text{Acc}_{\text{worst}}$).

		Unperturbed Metrics				Perturbed Metrics		
Benchmark	Prompt Type	Acc	Avg. δ_{TCB}	Avg. $\mathcal{V}_{\text{eff}}$	Avg. $z_k - z_{j^*}$	$\text{AccVar}_{\text{pert}}$	PDR (%)	$\text{Acc}_{\text{worst}}$
Very Confident Questions (VCQ Set)								
MMLU	Baseline	0.90	771.5	1.08	4.5	0.05	0.10	0.80
	Enhanced	**0.95**	**1025.2**	**1.03**	**6.0**	**0.02**	**0.03**	**0.90**
GSM8K	Baseline	0.85	2407.0	1.01	4.2	0.06	0.12	0.75
	Enhanced	**0.92**	**4410.8**	**1.05**	**5.8**	**0.03**	**0.05**	**0.85**
Ambiguous Questions (AQ Set)								
MMLU	Baseline	0.40	1983.0	1.01	1.5	0.15	30	0.15
	Enhanced	**0.70**	**2734.0**	**1.00**	**4.0**	**0.07**	**10**	**0.30**
GSM8K	Baseline	0.35	3412.8	1.04	1.2	0.18	35	0.10
	Enhanced	**0.65**	**6625.5**	**1.02**	**3.8**	**0.08**	**12**	**0.45**

correct yet brittle predictions; (2) **Inaccurate but Stable**: robustly wrong predictions; (3) **Confident but Unstable**: high confidence indicators (e.g., $P(\text{top1})$ or $z_k - z_{j^*}$) but low δ_{TCB}; (4) **Uncertain but Stable**: flatter distribution but a resilient underlying state. These are elaborated with examples in Appendix D.1 .

Case Study: Systematic Prompt Optimization Guided by δ_{TCB} **Objective:** To demonstrate using δ_{TCB} for guiding prompt engineering toward more robust solutions. **Method Summary:** We followed an iterative enhancement process on MMLU & GSM8K, detailed in Appendix C.4.2 . First, we ran baseline prompts over 3-5 random seeds to identify "Very Confident Questions" (VCQ; high and stable accuracy) and "Ambiguous Questions" (AQ; low/unstable accuracy or low δ_{TCB}). Second, for AQ sets, we performed targeted prompt engineering, systematically refining components like ICL examples and instructional phrasing to co-optimize for both accuracy and δ_{TCB}. Finally, we evaluated the enhanced prompts on unperturbed and perturbed data to measure gains in performance and robustness.

Results: Table 3 (illustrative of observed trends) shows that δ_{TCB}-guided enhanced prompts achieve higher Acc and **significantly higher mean δ_{TCB}** (e.g., for the MMLU AQ set, from 1983.0 to **2734.0**). More critically, they exhibit **superior robustness to perturbations**, exemplified by lower Performance Drop Rate (PDR) (e.g., MMLU AQ set PDR: $30\% \to \mathbf{10\%}$) and higher worst-case accuracy ($\text{Acc}_{\text{worst}}$) (e.g., MMLU AQ set $\text{Acc}_{\text{worst}}$: $15\% \to \mathbf{30\%}$). This underscores that co-optimizing for δ_{TCB} yields more dependable LLM performance, particularly under minor contextual shifts. It is noteworthy that even for the Ambiguous Questions (AQ) set, the average $\mathcal{V}_{\text{eff}}$ remains low (cf. Table 3), suggesting that ambiguity in correctness does not necessarily correspond to low model confidence; the model can be confidently wrong.

To benchmark δ_{TCB}-guided optimization, we compared it against a baseline strategy of perplexity-guided selection, where prompts are chosen to minimize the negative log-probability of the target answer. As shown in Table 4 , while perplexity-guidance improves accuracy, co-optimizing for δ_{TCB} yields more robust solutions with higher stability and better worst-case performance under perturbation.

Table 4: **Comparison of Prompt Optimization Strategies.** Co-optimizing for δ_{TCB} leads to more robust prompts compared to solely optimizing for perplexity (PPL), showing higher worst-case accuracy ($\text{Acc}_{\text{worst}}$) under perturbation.

Optimization Strategy	Avg. Acc	Avg. $\delta_{\text{TCB}} \uparrow$	Avg. PPL $\downarrow$	$\text{Acc}_{\text{worst}}$
Baseline Prompt	0.55	15.4	3.2	0.25
PPL-Guided	0.70	18.2	**1.9**	0.40
δ_{TCB}-Guided (Ours)	**0.72**	**35.8**	2.4	**0.65**

4.3.1 Deep Dive: Impact of Prompt Component Interactions on GSM8K

Objective: To showcase δ_{TCB}'s fine-grained diagnostic capabilities in dissecting prompt component effects. **Method Summary:** On GSM8K problem `gsm8k_811`, we systematically varied ICLs, instructions, and question phrasing from an accurate baseline. **Results:** Table 5 highlights counter-intuitive accuracy-stability trade-offs. For instance, simply adding a clarifying phrase ("7 days",

Table 5: **Impact of Prompt Interventions on Accuracy and Stability Metrics for a GSM8K question.** Striking trade-offs between Acc and δ_{TCB} emerge. For instance, adding "(7 days)" (Row 2) tanks Acc to 0% but **boosts δ_{TCB} significantly (8.20 → 46.97)**, indicating a stable but incorrect state. Row 7 shows an extreme case: Zero-Shot with a strong instruction results in 0% Acc but **astronomical $\delta_{\text{TCB}} \approx 49$k**, epitomizing a "confidently and extremely stably wrong" prediction. Metrics shown are Accuracy (Acc), Token Constraint Bound (δ_{TCB}), Effective Vocabulary Size ($\mathcal{V}_{\text{eff}}$), and top-2 logit margin ($z_k - z_{j^*}$).

Index	Intervention Description (`gsm8k_811`)	Acc (%)	$\delta_{\text{TCB}} \uparrow$	$\mathcal{V}_{\text{eff}} \downarrow$	$z_k - z_{j^*} \uparrow$
1	Baseline (New Algebraic ICLs, Original Question)	**100.0**	8.20	1.54	3.25
2	Clarified Q ("7 days") + New Alg. ICLs	0.00	**46.97**	1.04	5.23
3	Zero-shot CoT Instr. + Clarified Q + New Alg. ICLs	0.00	10.95	1.44	2.09
4	Role-Playing Instr. + Clarified Q + New Alg. ICLs	0.00	**62.14**	1.03	5.98
5	Algebraic Decomposition Instr. + Clarified Q + New Alg. ICLs	0.00	10.38	1.33	3.62
6	Hyper-Specific ICL + Alg. Decomp. Instr. + Clarified Q	0.00	**103.87**	1.02	5.55
7	Zero-Shot (No ICLs) + Alg. Decomp. Instr. + Clarified Q	0.00	**49450.23**	**1.00**	**11.29**
8	Formal Language Instr. + Clarified Q + New Alg. ICLs	0.00	**58.28**	1.04	5.32

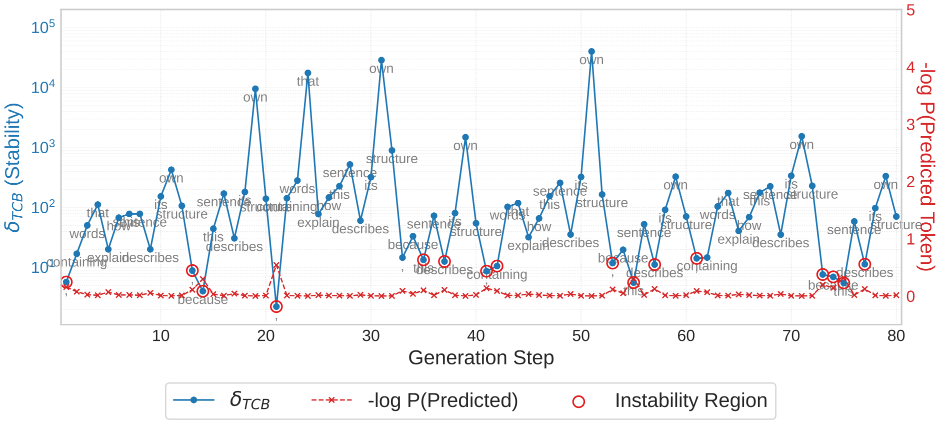


Figure 4: δ_{TCB} **dynamics vs.** P(**2nd best**) **during potentially repetitive generation.** Plot shows δ_{TCB} (blue, left y-axis) and P(2nd best) (green, right y-axis) versus generation step for LLaMA-3.1–8B. **Sharp dips in** δ_{TCB} (e.g., around steps 5-10, 20-25) often correlate with **spikes in** P(**2nd best**), indicating transient local instability not captured by average sequence PPL. Later, high, stable δ_{TCB} (e.g., steps 30+) can characterize a degenerate loop, showing robust commitment to the repetitive pattern.

Row 2) decimated accuracy (100% → 0%) but **boosted δ_{TCB} substantially (from 8.20 to 46.97)**, inducing a stable yet incorrect state. An even more extreme case is Row 7, where a zero-shot setup with a strong algebraic instruction yielded 0% accuracy but an **astronomical $\delta_{\text{TCB}} \approx 49$k** and perfect confidence ($\mathcal{V}_{\text{eff}} = 1.00$, $z_k - z_{j^*} = 11.29$). This epitomizes an extremely "confidently and stably wrong" prediction. These findings underscore δ_{TCB}'s capacity to uncover complex failure modes where models are robustly committed to erroneous reasoning pathsinsights that accuracy or standard confidence scores alone cannot provide.

5 Conclusion and Future Work

To address Large Language Model sensitivity to input context variations, this paper introduces the Token Constraint Bound (δ_{TCB}), a novel metric quantifying the local stability of next-token predictions against internal state perturbations. Intrinsically linked to the o_i^2-weighted geometric dispersion of output embeddings, δ_{TCB} offers a principled measure of an LLM's predictive commitment resilience. Our experiments demonstrate δ_{TCB}'s utility in assessing prompt effectiveness and its ability to uncover critical prediction instabilities missed by perplexity, thus providing a valuable complementary tool for analyzing and potentially enhancing LLM contextual robustness. While these findings are promising, our current investigation primarily uses a specific LLM and a focused set of scenarios. Future work should expand this research across a broader spectrum of models, varying scales, and diverse application contexts to validate and generalize the utility of δ_{TCB}. Exploring its application to understanding perturbations at intermediate layers could also yield deeper insights into representation robustness. Further investigation into δ_{TCB}'s role in model editing, fine-tuning, and robustness against more structured attacks remains an important avenue.

REFERENCES

Aryan Agrawal, Lisa Alazraki, Shahin Honarvar, and Marek Rei. Enhancing llm robustness to perturbed instructions: An empirical study. *arXiv preprint arXiv:2504.02733*, 2025.

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.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021.

Nurit Cohen-Inger, Yehonatan Elisha, Bracha Shapira, Lior Rokach, and Seffi Cohen. Forget what you know about llms evaluations-llms are like a chameleon. *arXiv preprint arXiv:2502.07445*, 2025.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda

Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Jiahui Geng, Zongxiong Chen, Yuandou Wang, Herbert Woisetschlaeger, Sonja Schimmler, Ruben Mayer, Zhiming Zhao, and Chunming Rong. A survey on dataset distillation: Approaches, applications and future directions. *arXiv preprint arXiv:2305.01975*, 2023.

Bahareh Harandizadeh, Abel Salinas, and Fred Morstatter. Risk and response in large language models: Evaluating key threat categories. *arXiv preprint arXiv:2403.14988*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint* *arXiv:2009.03300*, 2020.

David Herrera-Poyatos, Carlos Peláez-González, Cristina Zuheros, Andrés Herrera-Poyatos, Virilo Tejedor, Francisco Herrera, and Rosana Montes. An overview of model uncertainty and variability in llm-based sentiment analysis. challenges, mitigation strategies and the role of explainability. *arXiv preprint arXiv:2504.04462*, 2025.

Judy Hoffman, Daniel A Roberts, and Sho Yaida. Robust learning with jacobian regularization. *arXiv preprint arXiv:1908.02729*, 2019.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. *arXiv preprint arXiv:2104.08315*, 2021.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. Perplexitya measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1): S63–S63, 1977.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. Uncertainty quantification and confidence calibration in large language models: A survey. *arXiv preprint arXiv:2503.15850*, 2025.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*, 2021.

Huan Ma, Jingdong Chen, Guangyu Wang, and Changqing Zhang. Estimating llm uncertainty with logits. *arXiv preprint arXiv:2502.00290*, 2025.

Gary Marcus. The next decade in ai: four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*, 2020.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.

Fuseini Mumuni and Alhassan Mumuni. Explainable artificial intelligence (xai): from inherent explainability to large language models. *arXiv preprint arXiv:2501.09967*, 2025.

Roman Novak, Yasaman Bahri, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. *arXiv preprint arXiv:1802.08760*, 2018.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, ukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, ukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2023.

Yahao Pang, Xingyuan Wu, Xiaojin Zhang, Wei Chen, and Hai Jin. Fedeat: A robustness optimization framework for federated llms. *arXiv preprint arXiv:2502.11863*, 2025.

Amirhossein Razavi, Mina Soltangheis, Negar Arabzadeh, Sara Salamat, Morteza Zihayat, and Ebrahim Bagheri. Benchmarking prompt sensitivity in large language models. In *European Conference on Information Retrieval*, pp. 303–313. Springer, 2025.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*, 2023.

Ayush Singh, Navpreet Singh, and Shubham Vatsal. Robustness of llms to perturbations in text. *arXiv preprint arXiv:2407.08989*, 2024.

Petr Sychev, Andrey Goncharov, Daniil Vyazhev, Edvard Khalafyan, and Alexey Zaytsev. When an llm is apprehensive about its answers–and when its uncertainty is justified. *arXiv preprint arXiv:2503.01688*, 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Martin Chadwick, Gaurav Singh Tomar, Xavier Garcia, Evan Senter, Emanuel Taropa, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Yujing Zhang, Ravi Addanki, Antoine Miech, Annie Louis, Laurent El Shafey, Denis Teplyashin, Geoff Brown, Elliot Catt, Nithya Attaluri, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogoziska, Vitaly Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, Hanzhao Lin, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lui, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yong Cheng, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çalar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, YaGuang Li, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakievi, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Gamaleldin Elsayed, Ed Chi, Mahdis Mahdieh, Ian Tenney, Nan Hua, Ivan Petrychenko, Patrick Kane, Dylan Scandinaro, Rishub Jain, Jonathan Uesato, Romina Datta, Adam Sadovsky, Oskar Bunyan, Dominik Rabiej, Shimu Wu, John Zhang, Gautam Vasudevan, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Betty Chan, Pam G Rabinovitch, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Sahitya Potluri, Jane Park, Elnaz Davoodi, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Chris Gorgolewski, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Paul Suganthan, Evan Palmer, Geoffrey Irving, Edward Loper, Manaal Faruqui, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Michael Fink, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikoaj Rybiski, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marin Georgiev, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Minnie Lui, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Lam Nguyen Thiet, Daniel Andor, Pedro Valenzuela, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Sarmishta Velury, Sebastian Krause, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Tejasi Latkar, Mingyang Zhang, Quoc Le, Elena Allica Abellan, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Sid Lall, Ken Franko, Egor Filonov, Anna Bulanova, Rémi Leblond, Vikas Yadav, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Hao Zhou, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Jeremiah Liu, Mark Omernick, Colton Bishop, Chintu Kumar, Rachel Sterneck, Ryan Foley, Rohan Jain, Swaroop Mishra, Jiawei Xia, Taylor Bos, Geoffrey Cideron, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Petru Gurita, Hila Noga, Premal Shah, Daniel J. Mankowitz, Alex Polozov, Nate Kushman, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Anhad Mohananey, Matthieu Geist, Sidharth Mudgal, Sertan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Quan Yuan, Sumit Bagri, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Aliaksei Severyn, Jonathan Lai, Kathy Wu, Heng-Tze Cheng, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Mark Geller, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Andrei Sozanschi, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Abhimanyu Goyal, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Sabaer Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Tao Zhu, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane

Gu, Chenxi Pang, Dustin Tran, Yeqing Li, Nir Levine, Ariel Stolovich, Norbert Kalb, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Balaji Lakshminarayanan, Charlie Deck, Shyam Upadhyay, Hyo Lee, Mike Dusenberry, Zonglin Li, Xuezhi Wang, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Summer Yue, Sho Arora, Eric Malmi, Daniil Mirylenka, Qijun Tan, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Steven Zheng, Francesco Pongetti, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Ragha Kotikalapudi, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Chenkai Kuang, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Pei Sun, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Ishita Dasgupta, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Yuan Liu, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Pluciska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Ivo Penchev, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Adam Kurzrok, Lynette Webb, Sahil Dua, Dong Li, Preethi Lahoti, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Taylan Bilal, Evgenii Eltyshev, Daniel Balle, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Adams Yu, Christof Angermueller, Xiaowei Li, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Kevin Brooks, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Komal Jalan, Dinghua Li, Ginger Perng, Blake Hechtman, Parker Schuh, Milad Nasr, Mia Chen, Kieran Milan, Vladimir Mikulik, Trevor Strohman, Juliana Franco, Tim Green, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2023.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Yuqing Wang and Yun Zhao. Rupbench: Benchmarking reasoning under perturbations for robustness evaluation in large language models. *arXiv preprint arXiv:2406.11020*, 2024.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

Mozhi Zhang, Mianqiu Huang, Rundong Shi, Linsen Guo, Chong Peng, Peng Yan, Yaqian Zhou, and Xipeng Qiu. Calibrating the confidence of large language models by eliciting fidelity. *arXiv preprint arXiv:2404.02655*, 2024.

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.

CONTENTS

A LLM Usage Statement

LLMs were used solely as auxiliary tools for paper polishing. They did not contribute to the generation of research ideas, the design of experiments, the development of methodologies, data analysis, or any substantive aspects of the research. All scientific content, conceptual contributions, and experimental results are entirely the work of the authors. The authors take full responsibility for the contents of this paper.

B Related Work

Our work on the Token Constraint Bound (δ_{TCB}) builds upon and differentiates itself from several lines of research in LLM evaluation, robustness, and interpretability.

LLM Sensitivity and the Need for Robustness Metrics The pronounced sensitivity of LLMs to subtle input variations is well-documented. Studies have demonstrated substantial performance fluctuations arising from minor alterations in prompt phrasing (Razavi et al., 2025), example ordering and selection (Zhao et al., 2021; Lu et al., 2021), or even formatting details (Sclar et al., 2023). This brittleness (Marcus, 2020) highlights an urgent need for evaluation methods that go beyond aggregate task performance to assess the inherent stability of LLM predictions. Importantly, increased model scale, despite adherence to scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022), does not invariably confer enhanced robustness to these nuanced changes (Lu et al., 2021; Wei et al., 2023), further motivating metrics like δ_{TCB} that focus on local stability. The imperative for reliable deployment in critical applications (Weidinger et al., 2021; Herrera-Poyatos et al., 2025; Harandizadeh et al., 2024) makes understanding and quantifying such instabilities paramount.

Limitations of Conventional Evaluation Metrics Conventional metrics offer limited insight into local predictive robustness, a gap δ_{TCB} aims to address. Task accuracy, while a primary indicator, provides an aggregate view that can mask underlying prediction fragility due to contextual shifts (Zhao et al., 2021). Perplexity (Jelinek et al., 1977), though standard for assessing sequence likelihood (Liang et al., 2022; Holtzman et al., 2021), aggregates probabilities and can obscure local dynamics and the true stability of the internal state underpinning a prediction (Cohen-Inger et al., 2025). It may not reflect genuine model conviction, especially in cases of "surface form competition" (Holtzman et al., 2021) or when the softmax normalizes over poorly supported options. Emerging metrics targeting confidence and calibration (Zhang et al., 2024; Tian et al., 2023; Geng et al., 2023; Zhao et al., 2021) focus on aligning output probabilities with the likelihood of correctness (Tian et al., 2023). While valuable, these do not directly quantify the resilience of a specific prediction's dominant rank against perturbations in the model's internal representation $\boldsymbol{h}$ (Liu et al., 2025). A well-calibrated, high-confidence prediction might still stem from an internally fragile state. δ_{TCB} measures this internal representational stability directly.

Internal State Dynamics and Representational Stability Our work intersects with research exploring LLM internal representations and their stability. Efforts to enhance robustness by, for example, aligning hidden states of perturbed instructions with original ones (Agrawal et al., 2025), implicitly underscore the importance of stable internal configurations. δ_{TCB} provides a direct, quantitative measure of this local stability specifically for the next-token prediction mechanism, assessing the "safety margin" or integrity of the current predictive commitment arising from $\boldsymbol{h}$. This connects to broader goals in Explainable AI (XAI) that seek to understand internal model workings (Mumuni & Mumuni, 2025), where δ_{TCB} can pinpoint internal decision points of high or low resilience.

Prompt Engineering, In-Context Learning, and Stable State Induction δ_{TCB} is particularly relevant to analyzing the effectiveness of prompt engineering and In-Context Learning (Brown et al., 2020; Dong et al., 2022). Our central hypothesis posits that effective contextualization, such as through well-designed prompts or informative ICL examples, guides LLMs not only to correct answers but also to more *stable* internal states $\boldsymbol{h}$, reflected by higher δ_{TCB} values. This aligns with findings suggesting ICL's efficacy often stems from its ability to clarify task structure (Min et al., 2022). δ_{TCB} can thus serve as a quantitative tool to assess how effectively different contextual inputs induce robust internal commitments to a predictive path, complementing metrics focused on epistemic "knowledge strength" (Ma et al., 2025).

Output Embedding Geometry and Predictive Stability A key theoretical underpinning of δ_{TCB}, detailed in Section 3 , is its intrinsic link to the geometric dispersion of output token embeddings. While the general importance of embedding space properties for model robustness is acknowledged

(Pang et al., 2025), δ_{TCB} establishes a specific, analytical relationship between the geometry of the output embeddings (weighted by current prediction probabilities o_i^2) and the local stability of the next-token prediction against perturbations in $\boldsymbol{h}$. This provides a mechanistic, geometric interpretation of local prediction stability, moving beyond probability-based analyses alone.

Complementarity with Uncertainty Quantification (UQ) δ_{TCB} is positioned as complementary to, rather than a replacement for, existing Uncertainty Quantification (UQ) methods (Liu et al., 2025; Sychev et al., 2025). While UQ approaches typically aim to gauge a model's confidence or "knowledge strength" often based on output probabilities (e.g., entropy, prediction margins), δ_{TCB} specifically assesses the robustness of the *predictive mechanism* itself to internal state fluctuations. A prediction can exhibit high output probability (high confidence by UQ standards) yet originate from an unstable internal state (low δ_{TCB}), indicating a "confidently unstable" scenario. Conversely, a moderately confident prediction might be highly stable. δ_{TCB} thus offers a distinct perspective on reliability, focusing on the local integrity and resilience of the model's current predictive commitment rather than solely its expressed certainty.

Perturbation-Based Robustness in Neural Networks and LLMs Prior work on NN robustness has leveraged Jacobian norms and perturbations to quantify sensitivity and improve generalization. (Novak et al., 2018) empirically study sensitivity via the Frobenius norm of the input Jacobian, finding trained NNs more robust near training data and linking it to generalization gaps. Subsequent methods regularize this norm for adversarial robustness, e.g., (Hoffman et al., 2019) for classification margins. In LLMs, recent studies evaluate robustness to input perturbations like typos or rephrasing (Singh et al., 2024; Mumuni & Mumuni, 2025), with benchmarks like RUPBench (Wang & Zhao, 2024) showing larger models' resilience. δ_{TCB} extends these by deriving an exact closed-form for the softmax-linear Jacobian norm, linking it to output embedding geometry, and repurposing it for contextual stability in ICL and prompt engineering—focusing on hidden state perturbations rather than inputs.

C Experimental Details

C.1 Datasets and Task-Specific Setup

- **MMLU (Hendrycks et al., 2020):** For general validation (e.g., correlations in Table 1), questions were sampled from the ‘test‘ splits of subjects like ‘formal_logic‘, ‘philosophy‘, ‘abstract_algebra‘, ‘moral_scenarios‘, and ‘professional_law‘. The standard multiple-choice format (A, B, C, D) was used. The critical prediction point for δ_{TCB} calculation was immediately before the model generated the single token corresponding to its chosen letter (e.g., ’A’, ’B’, ’C’, or ’D’).
- **GSM8K (Cobbe et al., 2021):** For detailed intervention and prompt optimization studies, questions were sampled from the ‘test‘ set. These are grade-school math word problems requiring multi-step reasoning. The critical prediction point for δ_{TCB} was typically before the model generated the first token of the final numerical answer, after producing its chain-of-thought (CoT) reasoning. The final answer is usually identified by a pattern like "The final answer is `\boxed{X}`.".

C.2 Prompting Strategies and Examples

C.2.1 MMLU Prompts

Zero-Shot Multiple Choice:

```
Question: {question_text}
Options:
A) {option_A_text}
B) {option_B_text}
C) {option_C_text}
D) {option_D_text}
Answer:
```

The model is expected to complete with ’A’, ’B’, ’C’, or ’D’.

C.2.2 GSM8K Prompts (Baseline and Interventions)

Zero-Shot Chain-of-Thought (CoT):

```
Question: {question_text}
Let’s think step by step.
```

The model generates the reasoning steps and concludes with "The final answer is `\boxed{X}`."
Few-Shot Chain-of-Thought (CoT) ($k = 5$ baseline for GSM8K):

```
Question: {exemplar1_question_text}
Let’s think step by step.
{exemplar1_CoT_solution}
The final answer is \boxed{{exemplar1_answer}}.
###
Question: {exemplar2_question_text}
Let’s think step by step.
{exemplar2_CoT_solution}
The final answer is \boxed{{exemplar2_answer}}.
###
... (3 more exemplars) ...
###
Question: {current_question_text}
Let’s think step by step.
```

Exemplars were typically drawn from the GSM8K ‘train‘ set.
Interventions for GSM8K problem `gsm8k_811` (as in Table 5): The base question `gsm8k_811` is: "Felix earns $0.25 for each branch he trims from a tree. He trimmed branches from 12 trees. If he earned $60 in total, what is the average number of branches he trimmed per tree?" (Correct Answer: 20)

- **Clarified Q ("7 days")**: The original question text was appended with: "(Felix works 7 days a week)." This clarification is irrelevant and misleading for this specific problem.
- **ICL Variations**:
  - *New Algebraic ICLs (Baseline ICLs for Table 12):* Exemplars were selected/written to emphasize setting up and solving algebraic equations (e.g., using variables like x, y).
  - *Hyper-Specific ICL:* An ICL example was crafted to be structurally almost identical to `gsm8k_811` (e.g., "John earns $X per item. He processed Y items from Z batches. If he earned $Total, what is the average items per batch?"), but with different numbers and context.
- **Instructional Prefixes**: These were typically inserted directly before "Let’s think step by step." in a few-shot setup, or as the main instruction in a zero-shot setup.
  - *Zero-shot CoT Instr. (implied for Col 3 of Table 12 in context):* Simply "Let’s think step by step." as the primary instruction for the new question.
  - *Role-Playing Instr.:* "You are a brilliant mathematician. Solve the following problem by showing your detailed work."
  - *Algebraic Decomposition Instr.:* "Decompose this problem algebraically. Define variables, set up equations, and solve them step-by-step to find the final answer."
  - *Formal Language Instr.:* "Use precise mathematical language and formal notation in your solution. Ensure each step is clearly justified."

For Table 5, "New Alg. ICLs" were used unless "Hyper-Specific ICL" or "No ICLs" (Zero-Shot) is specified.

C.3 δ_{TCB} Calculation and Parameters

The Token Constraint Bound $\delta_{\mathrm{TCB}}(\boldsymbol{h})$ was computed using Eq. (10) from the main text:

$$\delta_{\mathrm{TCB}}(\boldsymbol{h}) = \frac{\epsilon}{\sqrt{\sum_{i=1}^{\mathcal{V}} o_i^2 \left\|\boldsymbol{w}_i - \boldsymbol{\mu}_{\boldsymbol{w}}(\boldsymbol{h})\right\|_2^2}}$$

The tolerance parameter ϵ was set to 1.0 for all experiments. This choice normalizes δ_{TCB} such that it represents the inverse of the Jacobian’s Frobenius norm (scaled by o_i^2-weighted embedding variance). As stated in the main text, relative changes and comparative values of δ_{TCB} are generally more informative than its absolute magnitude, which depends on ϵ. The hidden state $\boldsymbol{h}$ was taken from the output of the final transformer layer, just before the unembedding layer.

C.4 Details of Specific Experiments

C.4.1 Geometry Sensitivity Simulation

For each input prompt: 1. The original hidden state $\boldsymbol{h}_{\text{orig}}$ and output probability distribution $\boldsymbol{o}_{\text{orig}}$ were obtained. $\delta_{\text{TCB}}(\boldsymbol{h}_{\text{orig}}, \mathbf{W}_{\text{orig}})$ was calculated. 2. The top $K = 10$ competitor tokens (tokens $j \neq$ top1 with the highest o_j) were identified. 3. To create $\mathbf{W}_{\text{cluster}}$: For each competitor embedding $\boldsymbol{w}_j$ ($j \in K$ competitors), its new position was $\boldsymbol{w}_j^{\text{cluster}} = \boldsymbol{w}_j^{\text{orig}} + \alpha(\boldsymbol{w}_{\text{top1}}^{\text{orig}} - \boldsymbol{w}_j^{\text{orig}})$, where $\alpha = 0.5$ (moving it halfway towards the top-1 token's embedding). Embeddings of other tokens remained unchanged. $\delta_{\text{TCB}}(\boldsymbol{h}_{\text{orig}}, \mathbf{W}_{\text{cluster}})$ was calculated using the original $\boldsymbol{h}_{\text{orig}}$ and $\boldsymbol{o}_{\text{orig}}$. 4. To create $\mathbf{W}_{\text{disperse}}$: For each competitor embedding $\boldsymbol{w}_j$, its new position was $\boldsymbol{w}_j^{\text{disperse}} = \boldsymbol{w}_j^{\text{orig}} - \beta(\boldsymbol{w}_{\text{top1}}^{\text{orig}} - \boldsymbol{w}_j^{\text{orig}})$, where $\beta = 0.5$ (moving it away from the top-1 token's embedding along the same line, by half the original distance). $\delta_{\text{TCB}}(\boldsymbol{h}_{\text{orig}}, \mathbf{W}_{\text{disperse}})$ was calculated. The hypothesis $\delta_{\text{TCB}}(\mathbf{W}_{\text{cluster}}) > \delta_{\text{TCB}}(\mathbf{W}_{\text{orig}}) > \delta_{\text{TCB}}(\mathbf{W}_{\text{disperse}})$ was then checked.

C.4.2 Systematic Prompt Optimization Protocol

1. Baseline Characterization: A set of questions (e.g., 50-100 from MMLU/GSM8K) was run with a baseline prompt (e.g., zero-shot for MMLU, 5-shot CoT for GSM8K) across multiple random seeds (e.g., $S = 3$ or $S = 5$) for ICL selection or minor phrasing variants to get Acc statistics. Metrics collected per question: Mean Accuracy, Accuracy Variance (across seeds), Mean δ_{TCB} (at critical token, averaged over seeds if multiple correct paths), δ_{TCB} Variance. **VCQ (Very Confident Questions)** were defined as those with, e.g., Mean Acc ≥ 0.9, Acc Var ≤ 0.05, Mean $\delta_{\text{TCB}} > T_H$, δ_{TCB} Var $< V_L$. **AQ (Ambiguous Questions)** were defined as those with, e.g., Mean Acc < 0.6, or Acc Var > 0.15, or Mean $\delta_{\text{TCB}} < T_L$, or δ_{TCB} Var $> V_H$. Thresholds (T_H, T_L, V_L, V_H) were set empirically based on observed distributions.

2.Targeted Prompt Enhancement: For AQ questions, prompt engineering efforts focused on increasing both Acc and δ_{TCB}. Techniques included: Refining ICL examples (e.g., ensuring CoT steps are clearer, more analogous to the target problem structure, varying reasoning styles). Modifying instructional phrases (e.g., adding "Be very careful with calculations," "Explain your reasoning clearly"). Switching prompting strategy (e.g., from 5-shot to 2-shot with very high-quality examples, or adding a self-reflection step). For MMLU, this could involve adding a directive like "Choose the best option and explain why." For VCQ questions, efforts might focus on further increasing δ_{TCB} if it wasn't already maximal, or ensuring robustness (see below).

2.Evaluation (Unperturbed and Perturbed): Enhanced prompts were evaluated on the same metrics. Robustness was tested using perturbed inputs (see Appendix C.4.3). Metrics like Accuracy Variance on perturbed inputs ($\text{AccVar}_{\text{pert}}$), Performance Drop Rate (PDR), and Worst-Case Accuracy ($\text{Acc}_{\text{worst}}$) under perturbation were key.

C.4.3 Perturbation Strategies for Robustness Evaluation

Perturbations were designed to be plausible minor variations:

- **Syntactic Perturbations (applied to the question text):**
  - Paraphrasing: Using a pre-trained paraphrasing model (e.g., T5-based) to generate 2-3 variants of the question.
  - Reordering: For questions with multiple clauses/conditions, reordering them if semantically permissible.
  - Synonym Replacement: Replacing 1-2 keywords with close synonyms.
- **Semantic Perturbations (primarily for ICL / Few-Shot setups):**
  - ICL Example Reordering: Changing the order of the few-shot examples in the prompt.
  - ICL Example Replacement: Replacing one of the k examples with another valid but perhaps slightly less similar or slightly lower-quality example from the training set.
  - Adding a minor distractor sentence to the prompt context.

The goal was not to make the task unsolvable but to test sensitivity to typical input variations.

C.4.4 Text Generation Dynamics Setup

The prompt used to induce potentially repetitive behavior was:

```
System: Repeat the following word exactly five times: 'banana'.
After repeating it five times, say 'Task finished.'.
```

```
User: Okay, I understand. I will now repeat the word 'banana' five times
and then say 'Task finished.'
Assistant:
```

The model (LLaMA-3.1–8B) was allowed to generate tokens greedily. δ_{TCB}, $P(\text{top1})$, $P(\text{2nd best})$, and $\mathcal{V}_{\text{eff}}$ were calculated at each token generation step. The figure typically shows dynamics if the model continues beyond the explicit instruction, e.g., by getting stuck in a loop of "banana" or "Task finished." or exhibiting other non-ideal behaviors. The "dips" in δ_{TCB} often occur at points where the model is less certain about continuing a pattern versus breaking out or switching to another token.

D Further Discussion and Examples

D.1 Elaboration on Accuracy-Stability Conflict Scenarios

The main text (Section 4.3) mentioned four key conflict scenarios. Here are more detailed conceptual examples:

1.Accurate but Unstable (High Acc, Low δ_{TCB}): *Scenario:* The model correctly answers a complex reasoning question, but its internal state $\boldsymbol{h}$ is near a decision boundary where a slight internal "wobble" could have led to a different (incorrect) reasoning path and answer. *Example (MMLU-like):* Q: "Which of these legal principles is most directly violated by ex post facto laws?" Correct Answer: "Nulla poena sine lege". Model predicts "Nulla poena sine lege" (Acc=1). However, its δ_{TCB} is low (e.g., 2.1). This might be because the embedding for another plausible but incorrect principle like "Stare decisis" is geometrically positioned such that a small perturbation to $\boldsymbol{h}$ could shift the logits sufficiently to make "Stare decisis" the top choice. The correct prediction is thus brittle.

2.Inaccurate but Stable (Low Acc, High δ_{TCB}): *Scenario:* The model makes a clear error, often due to a misinterpretation or flawed reasoning pattern, but it is very robustly committed to this error. *Example (GSM8K-like, from Table 5 , Row 2):* Felix problem with the misleading "7 days" clarification. The model incorrectly calculates the answer (Acc=0) but does so with high δ_{TCB} (46.97). It has latched onto a stable, but flawed, interpretation/procedure.

3.Confident but Unstable (High $P(\textbf{top1})/z_k - z_{j^*}$, but Low δ_{TCB}): *Scenario:* The output probability for the top token is high, and/or the logit margin to the next token is large, suggesting strong confidence. However, the internal state supporting this is not robust. *Example:* Prompt: "What is the primary ingredient in concrete?" Model predicts "Cement" with $P(\text{Cement}) = 0.95$ and a large logit margin $z_k - z_{j^*} = 5.0$. Superficially, this looks very confident. However, δ_{TCB} is low (e.g., 1.8). This could occur if: The embedding $\boldsymbol{w}_{\text{Cement}}$ is relatively far from the mean embedding $\boldsymbol{\mu}_{\boldsymbol{w}}(\boldsymbol{h})$ (making the $o^2_{\text{Cement}}\|\boldsymbol{w}_{\text{Cement}} - \boldsymbol{\mu}_{\boldsymbol{w}}(\boldsymbol{h})\|_2^2$ term large despite o^2_{Cement} being large too, if the distance is very large). Or, many other low-probability competitor embeddings $\{\boldsymbol{w}_j\}$ are clustered very tightly around $\boldsymbol{\mu}_{\boldsymbol{w}}(\boldsymbol{h})$, leading to many small $o_j^2\|\boldsymbol{w}_j - \boldsymbol{\mu}_{\boldsymbol{w}}(\boldsymbol{h})\|_2^2$ terms whose sum is significant, contributing to a large Jacobian norm. The high probability $P(\text{Cement})$ might hide an underlying geometric configuration that is sensitive to perturbation.

4.Uncertain but Stable (Low $P(\textbf{top1})/z_k - z_{j^*}$, High $\mathcal{V}_{\text{eff}}$, but High δ_{TCB}): *Scenario:* The model's output distribution is relatively flat, indicating uncertainty among several top choices. Yet, the internal state representing this uncertainty is stable. *Example:* Prompt: "Which of these is a common pet: A) Dog, B) Tiger, C) Whale, D) Ant". Model output: $P(\text{Dog}) = 0.5$, $P(\text{Ant}) = 0.3$ (perhaps due to "common"), $P(\text{Tiger}) = 0.1$, $P(\text{Whale}) = 0.1$. $\mathcal{V}_{\text{eff}}$ is relatively high. However, δ_{TCB} could be high if the embeddings $\boldsymbol{w}_{\text{Dog}}$ and $\boldsymbol{w}_{\text{Ant}}$ are very close to each other (and thus to $\boldsymbol{\mu}_{\boldsymbol{w}}(\boldsymbol{h})$ if they dominate), while $\boldsymbol{w}_{\text{Tiger}}$ and $\boldsymbol{w}_{\text{Whale}}$ are far away. The model is stably "stuck" deciding between "Dog" and "Ant", but it's not about to suddenly jump to "Tiger". The uncertainty itself has a stable geometric basis.

D.2 Detailed Interpretation of GSM8K Intervention Analysis

This provides a row-by-row interpretation of the results for GSM8K question `gsm8k_811` shown in Table 5 of the main paper.

- **Row 1: Baseline (New Algebraic ICLs, Original Question)** *Metrics:* Acc=100%, $\delta_{\text{TCB}} = 8.20$, $\mathcal{V}_{\text{eff}} = 1.54$, $z_k - z_{j^*} = 3.25$. *Interpretation:* The baseline prompt with algebraic ICLs successfully solves the problem. The stability ($\delta_{\text{TCB}} = 8.20$) is moderate, indicating a reasonably robust correct prediction but with potential for improvement. $\mathcal{V}_{\text{eff}}$ and $z_k - z_{j^*}$ reflect good confidence.

- **Row 2: Clarified Q ("7 days") + New Alg. ICLs** *Metrics:* Acc=0%, $\delta_{\text{TCB}} = 46.97$, $\mathcal{V}_{\text{eff}} = 1.04$, $z_k - z_{j^*} = 5.23$. *Interpretation:* Adding the misleading "7 days" clarification breaks accuracy completely. Crucially, δ_{TCB} *increases dramatically*. This indicates the model has latched onto an incorrect interpretation or reasoning path due to the "7 days" phrase, and this incorrect path is highly stable. The high confidence metrics ($\mathcal{V}_{\text{eff}} \approx 1, z_k - z_{j^*}$ high) support this: it's "confidently and stably wrong."
- **Row 3: Zero-shot CoT Instr. + Clarified Q + New Alg. ICLs** *Metrics:* Acc=0%, $\delta_{\text{TCB}} = 10.95$, $\mathcal{V}_{\text{eff}} = 1.44$, $z_k - z_{j^*} = 2.09$. *Interpretation:* The standard "Let's think step by step" instruction does not fix the error caused by "7 days". The stability ($\delta_{\text{TCB}} = 10.95$) is higher than the original baseline (Row 1) but much lower than the "stably wrong" state in Row 2. This suggests the CoT instruction interacts with the misleading clarification and ICLs to produce a state that is still incorrect but less internally committed/stable than Row 2.
- **Row 4: Role-Playing Instr. + Clarified Q + New Alg. ICLs** *Metrics:* Acc=0%, $\delta_{\text{TCB}} = 62.14$, $\mathcal{V}_{\text{eff}} = 1.03$, $z_k - z_{j^*} = 5.98$. *Interpretation:* Accuracy remains 0. The Role-Playing instruction ("You are a brilliant mathematician...") leads to an even higher δ_{TCB} than Row 2, suggesting this type of instruction might encourage the model to commit more strongly to a particular reasoning path, even if that path is flawed due to other elements like the "7 days" clarification. Again, very high confidence in the wrong answer.
- **Row 5: Algebraic Decomposition Instr. + Clarified Q + New Alg. ICLs** *Metrics:* Acc=0%, $\delta_{\text{TCB}} = 10.38$, $\mathcal{V}_{\text{eff}} = 1.33$, $z_k - z_{j^*} = 3.62$. *Interpretation:* Similar to Row 3, this more specific algebraic instruction fails to correct the error. The stability is low, suggesting a conflict: the instruction pushes for algebraic rigor, but the "7 days" phrase may prevent a coherent algebraic formulation of the (misinterpreted) problem, leading to an unstable incorrect state.
- **Row 6: Hyper-Specific ICL + Alg. Decomp. Instr. + Clarified Q** *Metrics:* Acc=0%, $\delta_{\text{TCB}} = 103.87$, $\mathcal{V}_{\text{eff}} = 1.02$, $z_k - z_{j^*} = 5.55$. *Interpretation:* Even with a perfectly analogous ICL example and an algebraic instruction, the "7 days" clarification persists in causing an error. The δ_{TCB} is extremely high. This implies the model rigidly follows the structure of the hyper-specific ICL. If the "7 days" clarification leads to a consistent misapplication of that structure (e.g., always multiplying by 7 at a certain step because it's done in the (misinterpreted) exemplar logic), the result is a very stable, confident, but incorrect prediction.
- **Row 7: Zero-Shot (No ICLs) + Alg. Decomp. Instr. + Clarified Q** *Metrics:* Acc=0%, $\delta_{\text{TCB}} \approx 49450$, $\mathcal{V}_{\text{eff}} = 1.00$, $z_k - z_{j^*} = 11.29$. *Interpretation:* This is the most striking result. Removing ICLs (which might have conflicting signals) and relying solely on the strong "Algebraic Decomposition" instruction in the presence of the "7 days" clarification leads to an astronomically high δ_{TCB} and perfect confidence metrics, yet 0% accuracy. The model is utterly convinced by its (flawed) algebraic decomposition of the misinterpreted problem. This is the epitome of a "confidently and extremely stably wrong" state.
- **Row 8: Formal Language Instr. + Clarified Q + New Alg. ICLs** *Metrics:* Acc=0%, $\delta_{\text{TCB}} = 58.28$, $\mathcal{V}_{\text{eff}} = 1.04$, $z_k - z_{j^*} = 5.32$. *Interpretation:* Similar to the Role-Playing instruction (Row 4), asking for formal language seems to stabilize the incorrect reasoning path derived from the "7 days" clarification and algebraic ICLs, leading to high δ_{TCB} and confidence in the error.

This detailed analysis demonstrates δ_{TCB}'s power in revealing how different prompt components interact to affect not just accuracy, but the internal stability and commitment of the model to its predictions, whether correct or incorrect.

D.3 Differentiating δ_{TCB} from Perplexity (PPL)

δ_{TCB} assesses internal state robustness, a quality distinct from PPL's measure of sequence likelihood or token-level prediction confidence. Key differentiators include: **Robustness of (Potentially Erroneous) Confident Predictions:** PPL (or low token probability) flags incorrect predictions (if ground truth is known) but doesn't indicate if the model is *stably committed* to an error. δ_{TCB} quantifies this "stable incorrectness" (e.g., Table 5, Row 7, where δ_{TCB} is exceptionally high for an incorrect answer). **Sensitivity to Output Embedding Geometry:** PPL is solely a function of probabilities and thus blind to the geometric arrangement of output embeddings, which critically impacts δ_{TCB} and the true local stability of the prediction (cf. Section 4.2 and Eq. (10)). **Detection of Local Instabilities During Text Generation:** Sequence-level PPL averages token likelihoods,

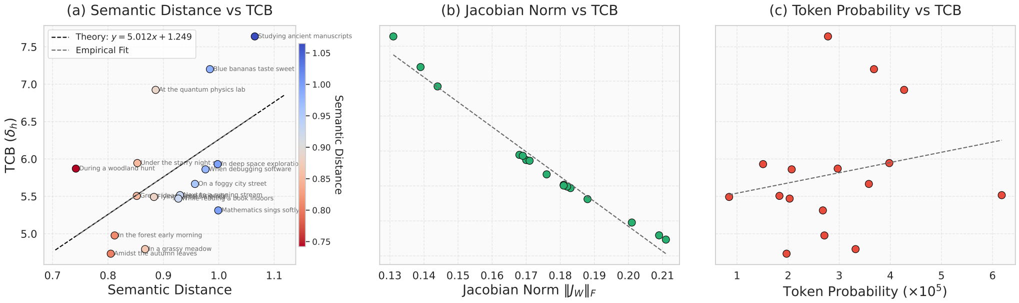


Figure 5: **Relationship between Prefix Semantics, TCB, and Model Internals.** Visual analysis based on 16 semantically varied prefixes targeting the same continuation. (a) δ_{TCB} versus semantic distance shows a strong positive linear correlation ($R^2 = 0.91$). Point colors map to semantic distance (see colorbar), and annotations identify prefixes. The empirical fit (gray dashed line) closely matches the data, aligning well with the theoretical prediction (black dashed line, Eq. (11)). (b) δ_{TCB} versus the Frobenius norm of the final layer Jacobian ($\|J_{\mathbf{W}}\|_F$) exhibits a moderate negative correlation. (c) δ_{TCB} versus the probability of the first token in the continuation shows a less distinct relationship for this set of prefixes.

potentially masking transient points of high internal instability that can precede errors or degenerate loops. Figure 4 illustrates δ_{TCB} capturing **sharp dips in** δ_{TCB} (often correlating with **spikes in** $P($**2nd best token**$)$). These crucial local dynamics, indicative of the model teetering between alternatives, are often smoothed over by aggregate PPL but are vital for understanding generation failures.

E Quantifying Semantic Context Influence.

Effective prompts often rely on precise semantic alignment. We investigated if δ_{TCB} quantitatively reflects this. Using 16 prefixes with varying semantic distances (cosine distance of embeddings) to a target prefix, we measured the δ_{TCB} of the first generated token. Figure 5 reveals a strong positive linear correlation ($R^2 = 0.91$) between semantic distance and δ_{TCB}. The empirical fit:

$$\delta_{\text{TCB}} = 5.012 \cdot \text{dist} + 1.249 \tag{11}$$

closely matches the data and theoretical predictions derived from sensitivity analysis. This shows δ_{TCB} provides a quantitative measure of how semantic (mis)alignment in the prompt impacts the stability of the resulting predictive state. Figure 5 also shows the expected negative correlation between δ_{TCB} and the Jacobian norm ($\|\mathbf{J}_{\mathbf{W}}\|_F$). This ability to quantify semantic influence further highlights δ_{TCB}'s utility for fine-grained prompt analysis and optimization.

F Statistical Derivation of δ_{TCB} Approximation

In this section, we provide a detailed derivation of a *statistical approximation* for the Token Constraint Bound (δ_{TCB}). This approach models the output weight matrix $\mathbf{W}$ as being drawn from a statistical ensemble and aims to approximate the TCB based on expected values, specifically targeting the root mean square (RMS) norm of the Jacobian. The resulting formula connects δ_{TCB} to model parameters like dimensionality and weight variance, as well as higher-order moments of the output probability distribution $\boldsymbol{o}$, offering potentially greater accuracy than simpler scaling laws, particularly when the output distribution is not diffuse (i.e., when $\mathcal{V}_{\text{eff}}$ is small). This statistical perspective contrasts with the exact, non-statistical expression derived in Section G, which applies deterministically to a specific instance of $\mathbf{W}$ and $\boldsymbol{o}$.
The quantity we aim to approximate is:

$$\delta_{\text{TCB}} = \frac{\epsilon}{\|\mathbf{J}_{\mathbf{W}}\|_F},$$

where the key terms involved are:

- δ_{TCB}: The Token Constraint Bound (Section 2.4), representing a distance in the hidden state space $\mathbf{R}^d$.

- $\epsilon > 0$: A fixed perturbation threshold, representing the target L_2 change in the output probability vector $\boldsymbol{o}$. This is typically treated as a dimensionless small value (e.g., 0.01).
- $\mathbf{J}_{\mathbf{W}} \in \mathbb{R}^{\mathcal{V} \times d}$: Jacobian of the softmax output probabilities $\boldsymbol{o}$ with respect to the hidden state $\boldsymbol{h}$. Its Frobenius norm $\|\mathbf{J}_{\mathbf{W}}\|_F$ has units of (probability change) / (hidden state unit).
- $\|\cdot\|_F$: The Frobenius norm of a matrix.
- $\boldsymbol{o} = (o_1, \ldots, o_{\mathcal{V}})^T \in \mathbf{R}^{\mathcal{V}}$: The vector of softmax output probabilities, $\sum_{i=1}^{\mathcal{V}} o_i = 1$, $o_i \geq 0$.
- $S_k = \sum_{i=1}^{\mathcal{V}} o_i^k$: The k-th moment sum of the output probabilities. Note $S_1 = 1$.
- $\mathcal{V}_{\text{eff}} = 1/S_2$: The effective vocabulary size, a measure of distribution flatness (strictly $\mathcal{V}_{\text{eff}}^{(2)}$).
- $\boldsymbol{h} \in \mathbf{R}^d$: The hidden state vector (embedding dimension d).
- $\mathbf{W} \in \mathbf{R}^{\mathcal{V} \times d}$: The output weight matrix (embedding matrix).
- σ^2: The assumed variance of the elements of $\mathbf{W}$ under the statistical model. This is a parameter of the theoretical model. When applied to a real model, it might represent an empirical estimate of the variance relevant to the specific forward pass or training state.
- $\mathcal{V}$: The size of the vocabulary.

Step 1: Recap the Jacobian Matrix $\mathbf{J}_{\mathbf{W}}$. As established in (4), the Jacobian is: Here, the matrix $\mathbf{M} \in \mathbf{R}^{\mathcal{V} \times \mathcal{V}}$ is symmetric and depends only on the output probability vector $\boldsymbol{o}$. It encapsulates how changes in the pre-softmax logits $\boldsymbol{z} = \mathbf{W}\boldsymbol{h}$ translate into changes in the post-softmax probabilities $\boldsymbol{o}$.

Step 2: Introduce Statistical Assumptions on W. The core of the statistical approach is to model the output weight matrix $\mathbf{W}$ not as a fixed, trained entity, but as a random matrix drawn from a simple ensemble. We make the following simplifying assumptions about its elements W_{jk}:

$$\mathbf{E}[W_{jk}] = 0 \quad \text{and} \quad \mathbf{E}[W_{jk}^2] = \sigma^2 \quad \text{(i.i.d.)}. \tag{12}$$

These imply $\mathbf{E}[W_{jk}W_{lm}] = \sigma^2 \delta_{jl}\delta_{km}$. The zero-mean assumption simplifies calculations. The constant variance σ^2 captures a typical scale.
Caveat 1 (Model Simplification): Realistically trained $\mathbf{W}$ matrices possess significant structure (e.g., semantic clusters, non-zero mean after layer normalization, varying variances per token, correlations between rows $\boldsymbol{w}_i, \boldsymbol{w}_j$). This i.i.d. model ignores such structure. Specifically, it implies $\mathbf{E}[\mathbf{W}\mathbf{W}^\top] = d\sigma^2 \mathbf{I}_{\mathcal{V}}$, assuming orthogonality between rows on average, which might not hold empirically. For enhanced accuracy, one might incorporate an empirical covariance term Σ_{emp} such that $\mathbf{W}\mathbf{W}^\top \approx d\sigma^2\mathbf{I} + \Sigma_{\text{emp}}$, though this complicates the derivation.
Note on σ^2: σ^2 should be interpreted either as a fixed parameter of the theoretical ensemble, or, when connecting to a real model, as an estimate of the empirical variance of weights relevant to the context (e.g., measured during the specific forward pass, possibly after normalization layers).

Step 3: Approximate the Squared Frobenius Norm via Expectation. We want to estimate $\|\mathbf{J}_{\mathbf{W}}\|_F^2 = \|\mathbf{M}\mathbf{W}\|_F^2$. Under the statistical model, $\|\mathbf{J}_{\mathbf{W}}\|_F^2$ is a random variable. We approximate it by its expectation $\mathbf{E}_{\mathbf{W}}[\|\mathbf{J}_{\mathbf{W}}\|_F^2]$.

$$\begin{aligned}
\mathbf{E}_{\mathbf{W}}[\|\mathbf{J}_{\mathbf{W}}\|_F^2] &= \mathbf{E}_{\mathbf{W}}\left[\text{Tr}(\mathbf{J}_{\mathbf{W}}\mathbf{J}_{\mathbf{W}}^\top)\right] \\
&= \mathbf{E}_{\mathbf{W}}\left[\text{Tr}(\mathbf{M}\mathbf{W}(\mathbf{M}\mathbf{W})^\top)\right] \\
&= \mathbf{E}_{\mathbf{W}}\left[\text{Tr}(\mathbf{M}\mathbf{W}\mathbf{W}^\top\mathbf{M}^\top)\right] \quad \text{(using } (AB)^\top = B^\top A^\top\text{)} \\
&= \mathbf{E}_{\mathbf{W}}\left[\text{Tr}(\mathbf{M}\mathbf{W}\mathbf{W}^\top\mathbf{M})\right] \quad \text{(since } \mathbf{M} \text{ is symmetric)} \qquad (13) \\
&= \mathbf{E}_{\mathbf{W}}\left[\sum_{i,j,l=1}^{\mathcal{V}} \sum_{k=1}^{d} M_{ij}W_{jk}W_{lk}M_{li}\right] \quad \text{(Trace expansion)} \qquad (14)
\end{aligned}$$

Here we introduce the first major approximation:
Approximation 1 (Decorrelation): The matrix elements $M_{ij} = \delta_{ij}o_i - o_io_j$ depend on $\boldsymbol{o} = \text{softmax}(\mathbf{W}\boldsymbol{h})$, which itself depends on $\mathbf{W}$. Therefore, $\mathbf{M}$ is correlated with $\mathbf{W}$. We make the strong approximation that this correlation can be ignored when computing the expectation involving quadratic terms of $\mathbf{W}$, effectively treating $\mathbf{M}$ as constant with respect to the expectation $\mathbf{E}_{\mathbf{W}}[\cdot]$.

$$\mathbf{E}_{\mathbf{W}}[\|\mathbf{J}_{\mathbf{W}}\|_F^2] \approx \text{Tr}\left(\mathbf{M}\mathbf{E}_{\mathbf{W}}[\mathbf{W}\mathbf{W}^\top]\mathbf{M}\right) \quad \textit{(Approximation 1 Applied)} \tag{15}$$

Caveat 2 (Decorrelation Risk): This approximation $\mathbf{E}[\mathbf{M}A\mathbf{M}] \approx \mathbf{M}\mathbf{E}[A]\mathbf{M}$ can introduce bias. The correlation is likely non-negligible if dimensions d or $\mathcal{V}$ are small, or if the distribution $\boldsymbol{o}$ is highly peaked (low $\mathcal{V}_{\text{eff}}$, where small changes in $\mathbf{W}$ affecting the top logits significantly alter $\mathbf{M}$). The error magnitude requires careful analysis (e.g., via perturbation theory or empirical validation in problematic regimes).

Now, we use the statistical property from (12). The (j,l)-th element of $\mathbf{E}_{\mathbf{W}}[\mathbf{W}\mathbf{W}^\top]$ is: $(\mathbf{E}_{\mathbf{W}}[\mathbf{W}\mathbf{W}^\top])_{jl} = \mathbf{E}_{\mathbf{W}}[(\mathbf{W}\mathbf{W}^\top)_{jl}] = \mathbf{E}_{\mathbf{W}}\left[\sum_{k=1}^{d} W_{jk}W_{lk}\right] = \sum_{k=1}^{d}\mathbf{E}_{\mathbf{W}}[W_{jk}W_{lk}] = \sum_{k=1}^{d}\sigma^2\delta_{jl} = d\sigma^2\delta_{jl}$. Therefore, $\mathbf{E}_{\mathbf{W}}[\mathbf{W}\mathbf{W}^\top] = d\sigma^2\mathbf{I}_{\mathcal{V}}$. Substituting this into (15):

$$\begin{aligned}\mathbf{E}_{\mathbf{W}}[\|\mathbf{J}_{\mathbf{W}}\|_F^2] &\approx \text{Tr}(\mathbf{M}(d\sigma^2\mathbf{I}_{\mathcal{V}})\mathbf{M}) \\ &= d\sigma^2\,\text{Tr}(\mathbf{M}^2) \quad (\text{Since } \mathbf{M}^T = \mathbf{M}) \\ &= d\sigma^2\|\mathbf{M}\|_F^2. \quad (\text{Definition of } \|\mathbf{M}\|_F^2 = \text{Tr}(\mathbf{M}\mathbf{M}^T))\end{aligned} \tag{16}$$

This indicates that the expected squared norm of the Jacobian is approximately proportional to the squared norm of the probability-dependent matrix $\mathbf{M}$, scaled by $d\sigma^2$. This connection relies critically on Approximation 1.

Step 4: Use the Exact Squared Frobenius Norm of $\mathbf{M}$. The calculation of $\|\mathbf{M}\|_F^2 = \text{Tr}(\mathbf{M}^2)$ is deterministic once $\boldsymbol{o}$ is known. As derived in Section G, the exact value is:

$$\|\mathbf{M}\|_F^2 = \sum_{i=1}^{\mathcal{V}} o_i^2(1-o_i)^2 + \sum_{i\neq j}(o_io_j)^2 \tag{17}$$

$$= \sum_i(o_i^2 - 2o_i^3 + o_i^4) + \sum_{i\neq j} o_i^2o_j^2 \tag{18}$$

$$= (S_2 - 2S_3 + S_4) + ((\sum_i o_i^2)(\sum_j o_j^2) - \sum_i (o_i^2)^2) \tag{19}$$

$$= S_2 - 2S_3 + S_4 + S_2^2 - S_4 \tag{20}$$

$$= S_2 - 2S_3 + S_2^2. \tag{21}$$

This expression relates the norm of $\mathbf{M}$ directly to the second (S_2) and third (S_3) moments of the output probability distribution. Note that this step involves no statistical approximation itself.

Step 5: Approximate the Jacobian Norm using RMS. We now introduce the second key approximation:

Approximation 2 (RMS Substitution): We approximate the actual Jacobian norm $\|\mathbf{J}_{\mathbf{W}}\|_F$ for a specific $\mathbf{W}$ by its root mean square (RMS) value under the statistical ensemble model:

$$\|\mathbf{J}_{\mathbf{W}}\|_F \approx \sqrt{\mathbf{E}_{\mathbf{W}}[\|\mathbf{J}_{\mathbf{W}}\|_F^2]}. \tag{22}$$

This relies on the assumption that the random variable $\|\mathbf{J}_{\mathbf{W}}\|_F$ concentrates around its RMS value (related to concentration of measure phenomena, plausible for large d or $\mathcal{V}$). However, it remains an approximation. By Jensen's inequality, $\sqrt{\mathbf{E}[X^2]} \geq \mathbf{E}[X]$, so the RMS value typically overestimates the expected norm. The magnitude of this overestimation (and the validity of concentration) depends on the variance of $\|\mathbf{J}_{\mathbf{W}}\|_F^2$, which might be large if the distribution of norms is heavy-tailed (e.g., due to specific weight structures or sparse outputs). More rigorous analysis might require bounding $\text{Var}[\|\mathbf{J}_{\mathbf{W}}\|_F]$ or using concentration inequalities (e.g., Hanson-Wright), which is beyond the scope of this derivation.

Substituting the result for the expected squared norm from (16) and the exact expression for $\|\mathbf{M}\|_F^2$ from (21) into (22):

$$\|\mathbf{J}_{\mathbf{W}}\|_F \approx \sqrt{d\sigma^2\|\mathbf{M}\|_F^2} = \sqrt{d\sigma^2(S_2 - 2S_3 + S_2^2)}. \tag{23}$$

This equation provides our refined statistical approximation for the Frobenius norm of the Jacobian.

Step 6: Refined Approximation for δ_{TCB}. Using the definition $\delta_{\text{TCB}} = \epsilon/\|\mathbf{J}_{\mathbf{W}}\|_F$ and substituting the RMS approximation for the norm from (23), we obtain the refined statistical approximation for the Token Constraint Bound:

$$\delta_{\text{TCB}} \approx \frac{\epsilon}{\sqrt{d\sigma^2(S_2 - 2S_3 + S_2^2)}}. \tag{24}$$

This is the main result of this section. It provides an estimate for δ_{TCB} based on the model's embedding dimension (d), the assumed variance of its output weights (σ^2), and the second (S_2) and third (S_3) moments of its current output probability distribution ($\boldsymbol{o}$).

Step 7: Connection to the Simpler $\mathcal{V}_{\text{eff}}$-based Approximation. A commonly cited simpler approximation for δ_{TCB} relates it directly to the effective vocabulary size $\mathcal{V}_{\text{eff}} = 1/S_2$, often presented as $\delta_{\text{TCB}} \approx \epsilon\sqrt{\mathcal{V}_{\text{eff}}/(d\sigma^2)}$. We see how this arises from (24) by introducing an *additional* approximation:
Approximation 3 (Diffuse Distribution): Assume $\boldsymbol{o}$ is sufficiently diffuse, meaning $p_{\max} = \max_i o_i$l1, corresponding to large effective vocabulary size, $\mathcal{V}_{\text{eff}} \gg 1$. Under this condition, higher moments S_k become small relative to lower moments. As argued in Section J ((51)), if o_i are roughly uniform over $\mathcal{V}_{\text{eff}}$ tokens ($o_i \sim 1/\mathcal{V}_{\text{eff}}$), then $S_3 \approx S_2/\mathcal{V}_{\text{eff}}$ and $S_2^2 \approx S_2/\mathcal{V}_{\text{eff}}$. Thus, for large $\mathcal{V}_{\text{eff}}$, $|-2S_3 + S_2^2|$lS_2.
This approximation $S_2 - 2S_3 + S_2^2 \approx S_2$ might fail if the distribution is not sufficiently "uniform-like" even if $\mathcal{V}_{\text{eff}}$ is large (e.g., heavy-tailed distributions where a few moderately high probabilities contribute significantly to S_3). Subject to this approximation:

$$S_2 - 2S_3 + S_2^2 \approx S_2 \quad (\text{Approximation 3: Diffuse distribution, } \mathcal{V}_{\text{eff}} \gg 1). \tag{25}$$

Substituting into (24):

$$\delta_{\text{TCB}} \approx \frac{\epsilon}{\sqrt{d\sigma^2 S_2}} = \frac{\epsilon}{\sqrt{d\sigma^2/\mathcal{V}_{\text{eff}}}} = \epsilon\sqrt{\frac{\mathcal{V}_{\text{eff}}}{d\sigma^2}}.$$

This shows the simpler $\mathcal{V}_{\text{eff}}$-based formula is a special case relying on both the statistical model and the diffuse distribution assumption.

G Exact, Non-Statistical Expression for $\|\mathbf{J}_{\mathbf{W}}\|_F$ and Its Relation to Weighted Variance

This section presents the derivation of the exact mathematical expression for the squared Frobenius norm $\|\mathbf{J}_{\mathbf{W}}\|_F^2$ for a *specific*, given output weight matrix $\mathbf{W} \in \mathbf{R}^{\mathcal{V} \times d}$ and hidden state $\boldsymbol{h} \in \mathbf{R}^d$ (which together determine a specific probability vector $\boldsymbol{o} = \operatorname{softmax}(\mathbf{W}\boldsymbol{h})$). This derivation is purely algebraic and deterministic; it does not rely on any statistical assumptions about $\mathbf{W}$ being drawn from a random ensemble. We establish the correct formula and clarify its relationship to, but distinctness from, the trace of the probability-weighted covariance matrix of the embeddings.
Notation Recap:

Step 1: Squared Frobenius Norm as Sum of Squared Row Norms. The squared Frobenius norm is the sum of the squared Euclidean norms of its rows. Let $\mathbf{J}_{\mathbf{W}}(i,:) \in \mathbb{R}^{1 \times d}$ denote the i-th row of the Jacobian $\mathbf{J}_{\mathbf{W}}$.

$$\|\mathbf{J}_{\mathbf{W}}\|_F^2 = \sum_{i=1}^{\mathcal{V}} \|\mathbf{J}_{\mathbf{W}}(i,:)\|_2^2. \tag{26}$$

Step 2: Derive the Expression for a Jacobian Row. We need the components of the i-th row, $\mathbf{J}_{\mathbf{W}}(i,:) = [(\mathbf{J}_{\mathbf{W}})_{i1}, \dots, (\mathbf{J}_{\mathbf{W}})_{id}]$. Recall $\mathbf{J}_{\mathbf{W}} = \boldsymbol{A}(\boldsymbol{o})\mathbf{W}$. The (i,k)-th element is:

$$\begin{aligned}
(\mathbf{J}_{\mathbf{W}})_{ik} &= \sum_{j=1}^{\mathcal{V}} A_{ij} W_{jk} \\
&= \sum_{j=1}^{\mathcal{V}} (\delta_{ij} o_i - o_i o_j) W_{jk} \quad \text{(Definition of } A_{ij}\text{)} \\
&= (o_i W_{ik} - o_i^2 W_{ik}) + \sum_{j \neq i} (-o_i o_j) W_{jk} \quad \text{(Splitting sum: } j = i \text{ and } j \neq i\text{)} \\
&= o_i W_{ik} - o_i \left(o_i W_{ik} + \sum_{j \neq i} o_j W_{jk} \right) \\
&= o_i W_{ik} - o_i \left(\sum_{j=1}^{\mathcal{V}} o_j W_{jk} \right) \quad \text{(Recombining sum)} \\
&= o_i (W_{ik} - (\sum_{j=1}^{\mathcal{V}} o_j W_{jk})) \\
&= o_i ((\boldsymbol{w}_i)_k - (\boldsymbol{\mu}_{\boldsymbol{w}})_k) = o_i (\boldsymbol{w}_i - \boldsymbol{\mu}_{\boldsymbol{w}})_k
\end{aligned} \tag{27}$$

Here, $W_{jk} = (\boldsymbol{w}_j)_k$ is the k-th component of embedding $\boldsymbol{w}_j$, and $(\sum_{j=1}^{\mathcal{V}} o_j W_{jk})$ is the k-th component of the mean embedding $\boldsymbol{\mu}_{\boldsymbol{w}}$. Thus, the entire i-th row vector (transposed to match $\boldsymbol{w}_i, \boldsymbol{\mu}_{\boldsymbol{w}}$ as column vectors) is:

$$\mathbf{J}_{\mathbf{W}}(i,:)^\top = o_i (\boldsymbol{w}_i - \boldsymbol{\mu}_{\boldsymbol{w}}). \tag{28}$$

This shows that the i-th row of the Jacobian represents the deviation of the i-th embedding from the mean embedding, scaled by the probability o_i.

Step 3: Substitute Row Norm back into Frobenius Norm Definition. Using the row expression (28) in the definition (26):

$$\begin{aligned}
\|\mathbf{J}_{\mathbf{W}}\|_F^2 &= \sum_{i=1}^{\mathcal{V}} \|\mathbf{J}_{\mathbf{W}}(i,:)\|_2^2 \\
&= \sum_{i=1}^{\mathcal{V}} \|o_i (\boldsymbol{w}_i - \boldsymbol{\mu}_{\boldsymbol{w}})^\top\|_2^2 \\
&= \sum_{i=1}^{\mathcal{V}} o_i^2 \|\boldsymbol{w}_i - \boldsymbol{\mu}_{\boldsymbol{w}}\|_2^2 \quad \text{(Since } o_i \text{ is a scalar).}
\end{aligned} \tag{29}$$

This is the exact, non-statistical expression for the squared Frobenius norm of the Jacobian matrix.

$$\|\mathbf{J}_{\mathbf{W}}(\boldsymbol{h})\|_F^2 = \sum_{i=1}^{\mathcal{V}} o_i^2 \|\boldsymbol{w}_i - \boldsymbol{\mu}_{\boldsymbol{w}}\|_2^2 \tag{30}$$

Step 4: Introduce the Trace of the Weighted Covariance Matrix. A related quantity often considered is the trace of the probability-weighted covariance matrix of the embedding vectors: Its trace represents the total variance of the embedding vectors, weighted by the probability distribution

$\boldsymbol{o}$:

$$\begin{aligned}\operatorname{Tr}[\operatorname{Cov}_{\boldsymbol{o}}(\boldsymbol{w})] &= \operatorname{Tr}\left[\sum_{i=1}^{\mathcal{V}} o_i(\boldsymbol{w}_i-\boldsymbol{\mu}_{\boldsymbol{w}})(\boldsymbol{w}_i-\boldsymbol{\mu}_{\boldsymbol{w}})^\top\right] \\ &= \sum_{i=1}^{\mathcal{V}} o_i \operatorname{Tr}\left[(\boldsymbol{w}_i-\boldsymbol{\mu}_{\boldsymbol{w}})(\boldsymbol{w}_i-\boldsymbol{\mu}_{\boldsymbol{w}})^\top\right] \quad \text{(Linearity of trace)} \\ &= \sum_{i=1}^{\mathcal{V}} o_i(\boldsymbol{w}_i-\boldsymbol{\mu}_{\boldsymbol{w}})^\top(\boldsymbol{w}_i-\boldsymbol{\mu}_{\boldsymbol{w}}) \quad \text{(Using } \operatorname{Tr}(\boldsymbol{a}\boldsymbol{b}^\top)=\boldsymbol{b}^\top\boldsymbol{a}\text{)} \\ &= \sum_{i=1}^{\mathcal{V}} o_i\|\boldsymbol{w}_i-\boldsymbol{\mu}_{\boldsymbol{w}}\|_2^2. \end{aligned} \tag{31}$$

This can also be expressed using the variance identity:

$$\operatorname{Tr}[\operatorname{Cov}_{\boldsymbol{o}}(\boldsymbol{w})] = \mathbb{E}_{\boldsymbol{o}}\left[\|\boldsymbol{w}\|_2^2\right] - \|\boldsymbol{\mu}_{\boldsymbol{w}}\|_2^2 = \left(\sum_{i=1}^{\mathcal{V}} o_i\|\boldsymbol{w}_i\|_2^2\right) - \left\|\sum_{j=1}^{\mathcal{V}} o_j\boldsymbol{w}_j\right\|_2^2. \tag{32}$$

Step 5: Comparing the Jacobian Norm and the Covariance Trace. Let us compare the exact squared Jacobian norm (30) with the trace of the covariance matrix (31):

$$\|\mathbf{J_W}\|_F^2 = \sum_{i=1}^{\mathcal{V}} o_i^2\|\boldsymbol{w}_i-\boldsymbol{\mu}_{\boldsymbol{w}}\|_2^2 \tag{33}$$

$$\operatorname{Tr}[\operatorname{Cov}_{\boldsymbol{o}}(\boldsymbol{w})] = \sum_{i=1}^{\mathcal{V}} o_i\|\boldsymbol{w}_i-\boldsymbol{\mu}_{\boldsymbol{w}}\|_2^2 \tag{34}$$

Step 6: The Exact Expression for δ_{TCB}. Using the correct formula for the squared Jacobian norm (30), the exact Token Constraint Bound is:

$$\delta_{\mathrm{TCB}} = \frac{\epsilon}{\|\mathbf{J_W}\|_F} = \frac{\epsilon}{\sqrt{\sum_{i=1}^{\mathcal{V}} o_i^2\,\|\boldsymbol{w}_i-\boldsymbol{\mu}_{\boldsymbol{w}}\|_2^2}}. \tag{35}$$

This is the ground-truth value for δ_{TCB} given a specific $\mathbf{W}$ and $\boldsymbol{h}$ (determining $\boldsymbol{o}$).

H Relating the Statistical Approximation and the Exact Expression

This section elucidates the precise relationship between the *refined statistical approximation* for $\|\mathbf{J_W}\|_F$ (derived in Section F) and the *exact, non-statistical expression* for $\|\mathbf{J_W}\|_F$ (derived in Section G). Understanding this connection is crucial for interpreting the validity and limitations of the statistical approach and for appreciating why it can serve as a useful, interpretable model despite its simplifying assumptions.

Recap: The Two Formulas for $\|\mathbf{J_W}\|_F^2$.
1. Refined Statistical Approximation (Section F): This approach models $\mathbf{W}$ as a random matrix. The core result approximates the actual squared norm by its expected value under the statistical model:

$$\mathbf{E_W}[\|\mathbf{J_W}\|_F^2] \approx d\sigma^2\|\mathbf{M}\|_F^2 = d\sigma^2(S_2-2S_3+S_2^2), \tag{36}$$

where $\mathbf{M} = \operatorname{diag}(\boldsymbol{o}) - \boldsymbol{o}\boldsymbol{o}^\top$, σ^2 is the assumed variance of W_{jk}, d is the embedding dimension, and $S_k = \sum_i o_i^k$. The final TCB approximation (24) uses the root mean square (RMS) value derived from this expectation: $\|\mathbf{J_W}\|_F \approx \sqrt{\mathbf{E_W}[\|\mathbf{J_W}\|_F^2]}$.
2. Exact Expression (Section G): For a specific, given $\mathbf{W}$ and $\boldsymbol{o}$, the squared norm is calculated deterministically using the geometry of the embeddings and the probability distribution:

$$\|\mathbf{J_W}\|_F^2 = \operatorname{Tr}[\operatorname{Cov}_{\boldsymbol{o}}(\boldsymbol{w})] = \mathbb{E}_{\boldsymbol{o}}\left[\|\boldsymbol{w}\|_2^2\right] - \|\mathbb{E}_{\boldsymbol{o}}[\boldsymbol{w}]\|_2^2 = \sum_{i=1}^{\mathcal{V}} o_i^2\|\boldsymbol{w}_i-\boldsymbol{\mu}_{\boldsymbol{w}}\|_2^2, \tag{37}$$

where $\boldsymbol{w}_i$ is the i-th embedding vector (row of $\mathbf{W}$), $\boldsymbol{\mu}_{\boldsymbol{w}} = \mathbb{E}_{\boldsymbol{o}}[\boldsymbol{w}] = \sum_j o_j \boldsymbol{w}_j$, and $\mathrm{Tr}[\mathrm{Cov}_{\boldsymbol{o}}(\boldsymbol{w})]$ is the trace of the probability-weighted covariance matrix of the embeddings. The exact TCB (35) uses the square root of this value.

The Fundamental Connection: Expectation Bridge. The theoretical link between these two expressions lies in taking the expectation of the *exact* squared norm (37) over the statistical ensemble assumed for $\mathbf{W}$. When we apply the statistical assumptions ((12)) and the key decorrelation approximation (**Approximation 1** from Section F, see (15)) to the exact formula, we recover the core quantity from the statistical approximation:

$$\mathbf{E}_{\mathbf{W}}\left[\underbrace{\|\mathbf{J}_{\mathbf{W}}\|_F^2}_{\text{Exact value from (37)}}\right] = \mathbf{E}_{\mathbf{W}}\left[\mathrm{Tr}(\boldsymbol{A}(\boldsymbol{o})\mathbf{W}\mathbf{W}^\top \boldsymbol{A}(\boldsymbol{o}))\right] \approx \underbrace{d\sigma^2\|\mathbf{M}\|_F^2}_{\text{Statistical result (36)}} . \tag{38}$$

Let's briefly trace why this works. The calculation performed in **Step 3** of Section F (starting from $\mathbf{E}_{\mathbf{W}}[\mathrm{Tr}(\mathbf{M}\mathbf{W}\mathbf{W}^\top\mathbf{M})]$) effectively computes the expectation of the exact squared norm, represented in the trace form old. The crucial step is applying **Approximation 1**:

$$\mathbf{E}_{\mathbf{W}}[\mathrm{Tr}(\mathbf{M}\mathbf{W}\mathbf{W}^\top\mathbf{M})] \approx \mathrm{Tr}(\mathbf{M}\mathbf{E}_{\mathbf{W}}[\mathbf{W}\mathbf{W}^\top]\mathbf{M}).$$

This approximation treats $\mathbf{M} = \mathrm{diag}(\boldsymbol{o}) - \boldsymbol{o}\boldsymbol{o}^\top$ as fixed when taking the expectation over $\mathbf{W}$, even though $\boldsymbol{o}$ itself depends on $\mathbf{W}$ via the softmax. Using $\mathbf{E}_{\mathbf{W}}[\mathbf{W}\mathbf{W}^\top]_{ij} = \mathrm{Tr}(\mathbf{E}[\boldsymbol{w}_i\boldsymbol{w}_j^\top]) = \sum_k \mathbf{E}[W_{ik}W_{jk}] = \sum_k \sigma^2\delta_{ij}\delta_{kk} = d\sigma^2\delta_{ij}$, we get $\mathbf{E}_{\mathbf{W}}[\mathbf{W}\mathbf{W}^\top] = d\sigma^2\mathbf{I}_{\mathcal{V}}$. Substituting this yields:

$$\mathrm{Tr}(\mathbf{M}(d\sigma^2\mathbf{I}_{\mathcal{V}})\mathbf{M}) = d\sigma^2\,\mathrm{Tr}(\mathbf{M}^2) = d\sigma^2\|\mathbf{M}\|_F^2.$$

This confirms that $d\sigma^2\|\mathbf{M}\|_F^2 = d\sigma^2(S_2 - 2S_3 + S_2^2)$ is indeed the result of calculating the expected value of the exact squared norm under the statistical assumptions and the decorrelation approximation.

Role of Approximations Revisited. The difference between the final δ_{TCB} value predicted by the statistical approximation (24) and the exact value (35) for a specific model arises precisely from the approximations inherent in the statistical derivation:

Statistical Ensemble Model for W: The actual trained weight matrix $\mathbf{W}$ is treated as a typical realization from a simple statistical ensemble (i.i.d., zero-mean, variance σ^2 elements, see (12)). Real trained matrices have complex structure (correlations, non-zero mean, varying variances) not captured by this model. The approximation's accuracy depends on how well the ensemble captures the properties relevant to the norm calculation (specifically, the second moments involved in $\mathbf{W}\mathbf{W}^\top$).

Decorrelation (M and W): The calculation in (38) relies on treating the probability-dependent matrix $\mathbf{M}$ as approximately uncorrelated with the quadratic terms in $\mathbf{W}$ (like $\mathbf{W}\mathbf{W}^\top$) when taking the expectation $\mathbf{E}_{\mathbf{W}}$ (**Approximation 1**, (15)). This approximation ignores the feedback loop where $\mathbf{W}$ determines $\boldsymbol{o}$, which in turn determines $\mathbf{M}$. It may be justified by averaging effects in high dimensions ($d, \mathcal{V}$), but it introduces a deviation from the exact expectation.

RMS Approximation ($\|\mathbf{J}_{\mathbf{W}}\|_F \approx \sqrt{\mathbf{E}[\|\mathbf{J}_{\mathbf{W}}\|_F^2]}$): The final step in the statistical approximation replaces the actual norm $\|\mathbf{J}_{\mathbf{W}}\|_F$ for the specific $\mathbf{W}$ with the ensemble-averaged RMS value (**Approximation 2**, (22)). This assumes that the norm of the specific matrix is close to the average norm across the ensemble, relying on concentration of measure phenomena. While often plausible for high-dimensional random matrices, the actual norm can deviate from the average.

In summary, the refined statistical TCB approximation (24) estimates the TCB by replacing the exact Jacobian norm $\|\mathbf{J}_{\mathbf{W}}\|_F$ with $\sqrt{d\sigma^2(S_2 - 2S_3 + S_2^2)}$, which represents the approximate RMS norm expected under the simplified statistical model and decorrelation assumption.

I Derivation of the Exact Jacobian Norm Formula

In this appendix, we provide a detailed step-by-step derivation of the exact mathematical expression for the squared Frobenius norm of the output Jacobian matrix, $\|\mathbf{J}_{\mathbf{W}}(\boldsymbol{h})\|_F^2$. This derivation is performed for a *specific* instance of the output weight matrix $\mathbf{W} \in \mathbb{R}^{\mathcal{V}\times d}$ and the hidden state $\boldsymbol{h} \in \mathbb{R}^d$, which together determine the output probability vector $\boldsymbol{o} = \mathrm{softmax}(\mathbf{W}\boldsymbol{h})$. The derivation relies solely on the definitions of the Jacobian and the Frobenius norm, requiring no statistical assumptions about $\mathbf{W}$.

The final result relates $\|\mathbf{J_W}\|_F^2$ to a weighted sum of squared distances involving the output embedding vectors. We also clarify its relationship to the trace of the probability-weighted covariance matrix of the embeddings, a concept central to related analyses but distinct from the Jacobian norm itself.

Step 1: Definition of Squared Frobenius Norm via Rows. The squared Frobenius norm of any matrix is the sum of the squared Euclidean norms (L_2 norms) of its rows. Let $\mathbf{J_W}(i,:)$ denote the i-th row vector of the Jacobian matrix $\mathbf{J_W}$.

$$\|\mathbf{J_W}(\boldsymbol{h})\|_F^2 = \sum_{i=1}^{\mathcal{V}} \|\mathbf{J_W}(i,:)\|_2^2. \tag{39}$$

Step 2: Jacobian Definition and its Relation to Weights. Recall the Jacobian matrix is given by $\mathbf{J_W} = \frac{\partial \boldsymbol{o}}{\partial \boldsymbol{h}}$. Using the chain rule, we can write:

$$\mathbf{J_W} = \frac{\partial \boldsymbol{o}}{\partial \boldsymbol{z}} \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{h}} = (\mathrm{diag}(\boldsymbol{o}) - \boldsymbol{o}\boldsymbol{o}^\top)\mathbf{W} = \boldsymbol{A}(\boldsymbol{o})\mathbf{W}.$$

The element $(\mathbf{J_W})_{ik}$ represents $\frac{\partial o_i}{\partial h_k}$.

Step 3: Explicit Calculation of the i-th Row of the Jacobian. We aim to find the vector $\mathbf{J_W}(i,:) = [(\mathbf{J_W})_{i1}, (\mathbf{J_W})_{i2}, \ldots, (\mathbf{J_W})_{id}]$. The component $(\mathbf{J_W})_{ik}$ is the (i,k)-th element of the matrix product $\boldsymbol{A}(\boldsymbol{o})\mathbf{W}$:

$$\begin{aligned}
(\mathbf{J_W})_{ik} &= \sum_{j=1}^{\mathcal{V}} A_{ij} W_{jk} \\
&= \sum_{j=1}^{\mathcal{V}} (\delta_{ij} o_i - o_i o_j) W_{jk} \quad \text{(Definition of } A_{ij}\text{)} \\
&= (o_i W_{ik}) + \sum_{j \neq i} (-o_i o_j) W_{jk} \\
&= o_i W_{ik} - o_i \sum_{j \neq i} o_j W_{jk} \\
&= o_i W_{ik} - o_i \left(\sum_{j=1}^{\mathcal{V}} o_j W_{jk} - o_i W_{ik} \right) \quad \text{(Completing the sum over } j\text{)} \\
&= o_i W_{ik} - o_i \left(\sum_{j=1}^{\mathcal{V}} o_j W_{jk} \right) + o_i^2 W_{ik}
\end{aligned}$$

Let's restart the derivation for $(\mathbf{J_W})_{ik}$ more directly:

$$\begin{aligned}
(\mathbf{J_W})_{ik} &= \sum_{j=1}^{\mathcal{V}} A_{ij} W_{jk} \\
&= A_{ii} W_{ik} + \sum_{j \neq i} A_{ij} W_{jk} \quad \text{(Separating diagonal term)} \\
&= o_i (1 - o_i) W_{ik} + \sum_{j \neq i} (-o_i o_j) W_{jk} \quad \text{(Substituting } A_{ij} \text{ values)} \\
&= o_i W_{ik} - o_i^2 W_{ik} - \sum_{j \neq i} o_i o_j W_{jk} \\
&= o_i W_{ik} - o_i \left(o_i W_{ik} + \sum_{j \neq i} o_j W_{jk} \right) \\
&= o_i W_{ik} - o_i \left(\sum_{j=1}^{\mathcal{V}} o_j W_{jk} \right) \quad \text{(Recombining sum)}
\end{aligned} \tag{40}$$

Now, let's interpret the terms. W_{jk} is the k-th component of the j-th embedding vector $\boldsymbol{w}_j$. Let $w_{j,k}$ denote this component.

$$\sum_{j=1}^{\mathcal{V}} o_j W_{jk} = \sum_{j=1}^{\mathcal{V}} o_j w_{j,k} = \left(\sum_{j=1}^{\mathcal{V}} o_j \boldsymbol{w}_j\right)_k = (\boldsymbol{\mu_w})_k$$

where $(\boldsymbol{\mu_w})_k$ is the k-th component of the mean embedding vector $\boldsymbol{\mu_w}$. Substituting this back into (40):

$$(\mathbf{J_W})_{ik} = o_i W_{ik} - o_i(\boldsymbol{\mu_w})_k = o_i(W_{ik} - (\boldsymbol{\mu_w})_k) = o_i(\boldsymbol{w}_i - \boldsymbol{\mu_w})_k \tag{41}$$

This expression gives the k-th component of the i-th row of $\mathbf{J_W}$. Therefore, the i-th row vector itself is:

$$\mathbf{J_W}(i,:) = o_i(\boldsymbol{w}_i - \boldsymbol{\mu_w})^\top. \tag{42}$$

This shows that each row of the Jacobian is the deviation of the corresponding embedding vector $\boldsymbol{w}_i$ from the mean embedding $\boldsymbol{\mu_w}$, scaled by the probability o_i.

Step 4: Substitute Row Norm back into Frobenius Norm Definition. Now we substitute the expression for the i-th row (42) back into the definition of the squared Frobenius norm (39):

$$\begin{aligned}\|\mathbf{J_W}(\boldsymbol{h})\|_F^2 &= \sum_{i=1}^{\mathcal{V}} \|\mathbf{J_W}(i,:)\|_2^2 \\ &= \sum_{i=1}^{\mathcal{V}} \|o_i(\boldsymbol{w}_i - \boldsymbol{\mu_w})^\top\|_2^2 \\ &= \sum_{i=1}^{\mathcal{V}} o_i^2 \|\boldsymbol{w}_i - \boldsymbol{\mu_w}\|_2^2 \quad \text{(Since } o_i \text{ is a scalar).}\end{aligned} \tag{43}$$

This equation provides the exact, non-statistical expression for the squared Frobenius norm of the Jacobian. It is determined by the squared distances between each embedding vector $\boldsymbol{w}_i$ and the mean embedding $\boldsymbol{\mu_w}$, weighted by the square of the corresponding probability o_i^2.

$$\|\mathbf{J_W}(\boldsymbol{h})\|_F^2 = \sum_{i=1}^{\mathcal{V}} o_i^2 \|\boldsymbol{w}_i - \boldsymbol{\mu_w}\|_2^2 \tag{44}$$

Step 5: Relationship to the Trace of the Covariance Matrix. The trace of the probability-weighted covariance matrix of the embedding vectors is a closely related but distinct concept. The covariance matrix is defined as: Its trace is:

$$\begin{aligned}\mathrm{Tr}[\mathrm{Cov}_{\boldsymbol{o}}(\boldsymbol{w})] &= \mathrm{Tr}\left[\sum_{i=1}^{\mathcal{V}} o_i(\boldsymbol{w}_i - \boldsymbol{\mu_w})(\boldsymbol{w}_i - \boldsymbol{\mu_w})^\top\right] \\ &= \sum_{i=1}^{\mathcal{V}} o_i \,\mathrm{Tr}\left[(\boldsymbol{w}_i - \boldsymbol{\mu_w})(\boldsymbol{w}_i - \boldsymbol{\mu_w})^\top\right] \quad \text{(Linearity of trace)} \\ &= \sum_{i=1}^{\mathcal{V}} o_i(\boldsymbol{w}_i - \boldsymbol{\mu_w})^\top(\boldsymbol{w}_i - \boldsymbol{\mu_w}) \quad (\text{Using } \mathrm{Tr}(\boldsymbol{a}\boldsymbol{b}^\top) = \boldsymbol{b}^\top\boldsymbol{a}) \\ &= \sum_{i=1}^{\mathcal{V}} o_i\|\boldsymbol{w}_i - \boldsymbol{\mu_w}\|_2^2.\end{aligned} \tag{45}$$

As derived previously, this trace also equals the variance identity:

$$\mathrm{Tr}[\mathrm{Cov}_{\boldsymbol{o}}(\boldsymbol{w})] = \mathbb{E}_{\boldsymbol{o}}\left[\|\boldsymbol{w}\|_2^2\right] - \|\mathbb{E}_{\boldsymbol{o}}[\boldsymbol{w}]\|_2^2 = \left(\sum_{i=1}^{\mathcal{V}} o_i\|\boldsymbol{w}_i\|_2^2\right) - \left\|\sum_{j=1}^{\mathcal{V}} o_j\boldsymbol{w}_j\right\|_2^2. \tag{46}$$

This quantity, $\mathrm{Tr}[\mathrm{Cov}_{\boldsymbol{o}}(\boldsymbol{w})]$, represents the total variance of the embedding vectors, weighted by the probabilities o_i.

Step 6: Comparing Jacobian Norm and Covariance Trace. Comparing the derived exact squared Jacobian norm (43) with the trace of the covariance matrix (45):

$$\|\mathbf{J}_{\mathbf{W}}\|_F^2 = \sum_{i=1}^{\mathcal{V}} o_i^2 \|\boldsymbol{w}_i - \boldsymbol{\mu}_{\boldsymbol{w}}\|_2^2 \tag{47}$$

$$\mathrm{Tr}[\mathrm{Cov}_{\boldsymbol{o}}(\boldsymbol{w})] = \sum_{i=1}^{\mathcal{V}} o_i \|\boldsymbol{w}_i - \boldsymbol{\mu}_{\boldsymbol{w}}\|_2^2 \tag{48}$$

These two quantities are clearly different, differing by a factor of o_i in the weighting term inside the sum. While both measure aspects of the dispersion of the embedding vectors relative to their mean $\boldsymbol{\mu}_{\boldsymbol{w}}$, they are not mathematically identical. The Jacobian norm gives more weight (via o_i^2) to the deviation of high-probability embeddings from the mean.

J Derivation of Regime-Dependent TCB Behavior

In this appendix, we provide detailed mathematical derivations supporting the regime-dependent behavior of the Token Constraint Bound (δ_{TCB}), as discussed in Section 3 and observed empirically Table 1. We analyze the behavior of $\delta_{\mathrm{TCB}} = \epsilon / \|\mathbf{J}_{\mathbf{W}}(\boldsymbol{h})\|_F$ in two distinct regimes based on the flatness of the output probability distribution $\boldsymbol{o}$: high flatness (large effective vocabulary size, $\mathcal{V}_{\mathrm{eff}} \gg 1$) and low flatness (highly peaked distribution, $\mathcal{V}_{\mathrm{eff}} \approx 1$).

The key is to understand how the squared Frobenius norm of the Jacobian, $\|\mathbf{J}_{\mathbf{W}}(\boldsymbol{h})\|_F^2$, behaves in these limits. We will use different but related expressions for the norm depending on the regime: the statistical approximation for the high-$\mathcal{V}_{\mathrm{eff}}$ regime and the exact formula for the low-$\mathcal{V}_{\mathrm{eff}}$ regime.

Recap of Key Formulas:

1. **Exact Squared Norm (from Section I, Eq.** (43)**):**

$$\|\mathbf{J}_{\mathbf{W}}(\boldsymbol{h})\|_F^2 = \sum_{i=1}^{\mathcal{V}} o_i^2 \|\boldsymbol{w}_i - \boldsymbol{\mu}_{\boldsymbol{w}}\|_2^2 \tag{49}$$

   where $\boldsymbol{w}_i$ is the i-th embedding vector (row of $\mathbf{W}$), $\boldsymbol{o} = (o_1, \ldots, o_{\mathcal{V}})$ is the probability vector, and $\boldsymbol{\mu}_{\boldsymbol{w}} = \sum_{j=1}^{\mathcal{V}} o_j \boldsymbol{w}_j$ is the probability-weighted mean embedding.

2. **Refined Statistical Approximation (from Section F, Eq.** (23)**):**

$$\|\mathbf{J}_{\mathbf{W}}(\boldsymbol{h})\|_F \approx \sqrt{d\sigma^2(S_2 - 2S_3 + S_2^2)} \tag{50}$$

   where d is the hidden dimension, σ^2 is the assumed variance of weight elements W_{jk}, and $S_k = \sum_i o_i^k$ is the k-th moment sum ($\mathcal{V}_{\mathrm{eff}} = 1/S_2$). This approximation relies on modeling $\mathbf{W}$ statistically and approximating the norm by its RMS value.

J.1 Regime 1: High Flatness (Large $\mathcal{V}_{\mathrm{eff}} \gg 1$)

Assumptions. In this regime, the probability distribution $\boldsymbol{o}$ is spread relatively evenly across many tokens.

- $\mathcal{V}_{\mathrm{eff}} = 1/S_2$ is large.
- Consequently, the maximum probability $p_{\max} = \max_i o_i$ must be small ($p_{\max}$l1). Roughly, if probabilities are spread over $\sim \mathcal{V}_{\mathrm{eff}}$ tokens, then $o_i \sim 1/\mathcal{V}_{\mathrm{eff}}$.
- We use the refined statistical approximation (50), which implicitly assumes the statistical model for $\mathbf{W}$ (i.i.d. elements, zero mean, variance σ^2) is a reasonable proxy for average behavior.

Approximating the Jacobian Norm. We analyze the term $\|\mathbf{M}\|_F^2 = S_2 - 2S_3 + S_2^2$ within the statistical approximation (50). Since o_il1 for all i, higher powers of o_i are much smaller. Let's assess the magnitude of the terms relative to S_2:

- $S_2 = \sum_i o_i^2 = 1/\mathcal{V}_{\mathrm{eff}}$.
- $S_3 = \sum_i o_i^3 \le (\max_j o_j) \sum_i o_i^2 = p_{\max} S_2$. If $o_i \sim 1/\mathcal{V}_{\mathrm{eff}}$, then $S_3 \sim \mathcal{V}_{\mathrm{eff}} \times (1/\mathcal{V}_{\mathrm{eff}})^3 = 1/\mathcal{V}_{\mathrm{eff}}^2 = S_2/\mathcal{V}_{\mathrm{eff}}$.
- $S_2^2 = (1/\mathcal{V}_{\mathrm{eff}})^2 = S_2/\mathcal{V}_{\mathrm{eff}}$.

Thus, both S_3 and S_2^2 are smaller than S_2 by a factor of approximately $\mathcal{V}_{\text{eff}}$. Since we assume $\mathcal{V}_{\text{eff}} \gg 1$, the terms $-2S_3$ and $+S_2^2$ become negligible compared to S_2:

$$\|\mathbf{M}\|_F^2 = S_2 - 2S_3 + S_2^2 \approx S_2 \quad \text{(for large } \mathcal{V}_{\text{eff}}\text{)}. \tag{51}$$

This simplification corresponds to Approximation 3 discussed in Section F. Substituting this back into the statistical norm approximation (50):

$$\|\mathbf{J_W}\|_F \approx \sqrt{d\sigma^2(S_2)} \tag{52}$$

$$= \sqrt{\frac{d\sigma^2}{\mathcal{V}_{\text{eff}}}}. \tag{53}$$

This expression predicts that the Jacobian norm decreases as $\mathcal{V}_{\text{eff}}$ increases.

Behavior of δ_{TCB}. Using the definition $\delta_{\text{TCB}} = \epsilon / \|\mathbf{J_W}\|_F$ and the approximation (53):

$$\delta_{\text{TCB}} \approx \frac{\epsilon}{\sqrt{d\sigma^2/\mathcal{V}_{\text{eff}}}} = \epsilon\sqrt{\frac{\mathcal{V}_{\text{eff}}}{d\sigma^2}}. \tag{54}$$

Conclusion (High $\mathcal{V}_{\text{eff}}$): In the high-flatness regime, δ_{TCB} is predicted to be approximately proportional to the square root of the effective vocabulary size:

$$\delta_{\text{TCB}} \propto \sqrt{\mathcal{V}_{\text{eff}}} \quad \text{(for } \mathcal{V}_{\text{eff}} \gg 1\text{)}$$

This provides a clear mathematical basis for the strong positive correlation $r_{\delta,\mathcal{V}_{\text{eff}}} \approx 0.95$ observed empirically (Table 1) in diverse datasets where distributions are often flat.

Negligible Influence of Margin $z_k - z_{j^*}$. The margin $z_k - z_{j^*} = z_k - z_{j^*}$ between the logits of two specific tokens k (usually the top prediction) and j^* (usually the top competitor) directly affects o_k and o_{j^*}. In the high-$\mathcal{V}_{\text{eff}}$ regime, however, both o_k and o_{j^*} are typically small (e.g., $\sim 1/\mathcal{V}_{\text{eff}}$). Changes in $z_k - z_{j^*}$ primarily redistribute a small amount of probability mass between these two (and possibly nearby) tokens. The impact on the overall sum $S_2 = \sum o_i^2$, which aggregates contributions from many small probabilities, is minimal. Consequently, changes in $z_k - z_{j^*}$ have a very weak effect on $\|\mathbf{J_W}\|_F$ via (53), and therefore also on δ_{TCB}. This explains the near-zero correlation $r_{\delta,z_k-z_{j^*}}$ observed in diverse datasets (Table 1).

J.2 Regime 2: Low Flatness (Small $\mathcal{V}_{\text{eff}} \approx 1$)

Assumptions. In this regime, the probability distribution $\boldsymbol{o}$ is sharply peaked on a single token.

- $\mathcal{V}_{\text{eff}} \approx 1$. This occurs when one probability, say o_k, is close to 1.
- Let $o_k = 1 - \epsilon_s$, where $\epsilon_s = \sum_{j \neq k} o_j$ is a small positive quantity (ϵ_sl1).
- All other probabilities o_j ($j \neq k$) are very small, typically $o_j \sim O(\epsilon_s)$ or smaller.
- We use the exact expression for the norm (49) as it directly captures the influence of the specific dominant embedding $\boldsymbol{w}_k$ and its relation to competitors. Statistical averaging inherent in (50) is less appropriate here.

Approximating the Jacobian Norm. We analyze the exact sum $\|\mathbf{J_W}\|_F^2 = \sum_{i=1}^{\mathcal{V}} o_i^2 \|\boldsymbol{w}_i - \boldsymbol{\mu_w}\|_2^2$. First, approximate the mean embedding $\boldsymbol{\mu_w}$:

$$\boldsymbol{\mu_w} = \sum_{j=1}^{\mathcal{V}} o_j \boldsymbol{w}_j = o_k \boldsymbol{w}_k + \sum_{j \neq k} o_j \boldsymbol{w}_j \tag{55}$$

$$= (1 - \epsilon_s)\boldsymbol{w}_k + \sum_{j \neq k} o_j \boldsymbol{w}_j \tag{56}$$

$$= \boldsymbol{w}_k - \epsilon_s \boldsymbol{w}_k + \sum_{j \neq k} o_j \boldsymbol{w}_j \tag{57}$$

$$= \boldsymbol{w}_k + \sum_{j \neq k} o_j (\boldsymbol{w}_j - \boldsymbol{w}_k) \quad \text{(using } \epsilon_s = \sum_{j \neq k} o_j\text{)} \tag{58}$$

This shows $\boldsymbol{\mu_w}$ is close to $\boldsymbol{w}_k$, differing by terms of order $O(\epsilon_s)$.
Now, consider the terms in the sum for $\|\mathbf{J_W}\|_F^2$:

- **Term for $i = k$:** We need $\|\boldsymbol{w}_k - \boldsymbol{\mu}_{\boldsymbol{w}}\|_2^2$. From (58):

  $$\boldsymbol{w}_k - \boldsymbol{\mu}_{\boldsymbol{w}} = -\sum_{j \neq k} o_j(\boldsymbol{w}_j - \boldsymbol{w}_k)$$

  The squared norm is $\|\boldsymbol{w}_k - \boldsymbol{\mu}_{\boldsymbol{w}}\|_2^2 = \left\|\sum_{j \neq k} o_j(\boldsymbol{w}_j - \boldsymbol{w}_k)\right\|_2^2$. Since each $o_j = O(\epsilon_s)$ for $j \neq k$, this squared norm is $O(\epsilon_s^2)$. The contribution to the total sum is $o_k^2\|\boldsymbol{w}_k - \boldsymbol{\mu}_{\boldsymbol{w}}\|_2^2 \approx (1-\epsilon_s)^2 O(\epsilon_s^2) \approx O(\epsilon_s^2)$. This term is therefore negligible to the leading order.

- **Terms for $i \neq k$:** We need $\|\boldsymbol{w}_i - \boldsymbol{\mu}_{\boldsymbol{w}}\|_2^2$. Using (58):

  $$\begin{aligned}\boldsymbol{w}_i - \boldsymbol{\mu}_{\boldsymbol{w}} &= \boldsymbol{w}_i - \left(\boldsymbol{w}_k + \sum_{j \neq k} o_j(\boldsymbol{w}_j - \boldsymbol{w}_k)\right) \\ &= (\boldsymbol{w}_i - \boldsymbol{w}_k) - \sum_{j \neq k} o_j(\boldsymbol{w}_j - \boldsymbol{w}_k)\end{aligned}$$

  The squared norm is:

  $$\|\boldsymbol{w}_i - \boldsymbol{\mu}_{\boldsymbol{w}}\|_2^2 = \left\|(\boldsymbol{w}_i - \boldsymbol{w}_k) - \sum_{j \neq k} o_j(\boldsymbol{w}_j - \boldsymbol{w}_k)\right\|_2^2$$

  Expanding the square:

  $$\|\boldsymbol{w}_i - \boldsymbol{\mu}_{\boldsymbol{w}}\|_2^2 = \|\boldsymbol{w}_i - \boldsymbol{w}_k\|_2^2 - 2(\boldsymbol{w}_i - \boldsymbol{w}_k)^\top \left(\sum_{j \neq k} o_j(\boldsymbol{w}_j - \boldsymbol{w}_k)\right) + \left\|\sum_{j \neq k} o_j(\boldsymbol{w}_j - \boldsymbol{w}_k)\right\|_2^2$$

  The middle term is $O(\epsilon_s)$, and the last term is $O(\epsilon_s^2)$. Thus, to leading order:

  $$\|\boldsymbol{w}_i - \boldsymbol{\mu}_{\boldsymbol{w}}\|_2^2 \approx \|\boldsymbol{w}_i - \boldsymbol{w}_k\|_2^2 + O(\epsilon_s) \quad (\text{for } i \neq k)$$

  The contribution of the i-th term ($i \neq k$) to the total sum $\|\mathbf{J}_{\mathbf{W}}\|_F^2$ is $o_i^2\|\boldsymbol{w}_i - \boldsymbol{\mu}_{\boldsymbol{w}}\|_2^2$. Since $o_i^2 = O(\epsilon_s^2)$, the contribution is:

  $$o_i^2\|\boldsymbol{w}_i - \boldsymbol{\mu}_{\boldsymbol{w}}\|_2^2 \approx o_i^2(\|\boldsymbol{w}_i - \boldsymbol{w}_k\|_2^2 + O(\epsilon_s)) = o_i^2\|\boldsymbol{w}_i - \boldsymbol{w}_k\|_2^2 + O(\epsilon_s^3)$$

Summing the contributions for all i:

$$\begin{aligned}\|\mathbf{J}_{\mathbf{W}}\|_F^2 &= \underbrace{o_k^2\|\boldsymbol{w}_k - \boldsymbol{\mu}_{\boldsymbol{w}}\|_2^2}_{\approx O(\epsilon_s^2)} + \sum_{j \neq k} \underbrace{o_j^2\|\boldsymbol{w}_j - \boldsymbol{\mu}_{\boldsymbol{w}}\|_2^2}_{\approx o_j^2\|\boldsymbol{w}_j - \boldsymbol{w}_k\|_2^2 + O(\epsilon_s^3)} \\ &\approx \sum_{j \neq k} o_j^2\|\boldsymbol{w}_j - \boldsymbol{w}_k\|_2^2 \quad (\text{keeping leading order terms, } O(\epsilon_s^2))\end{aligned} \tag{59}$$

This approximation reveals that for highly peaked distributions, the squared Jacobian norm is dominated by the sum of squared distances between the dominant embedding $\boldsymbol{w}_k$ and competitor embeddings $\boldsymbol{w}_j$, weighted by the *square* of the competitors' small probabilities o_j^2.

Connecting to Margin $z_k - z_{j^*}$. The logit margin between the winning token k and any other token j is $z_k - z_{j^*} = z_k - z_j$. The probability o_j for $j \neq k$ can be approximated using the softmax definition when z_k is large compared to others:

$$o_j = \frac{e^{z_j}}{\sum_{l=1}^{\mathcal{V}} e^{z_l}} = \frac{e^{z_j}}{e^{z_k} + \sum_{l \neq k} e^{z_l}} \approx \frac{e^{z_j}}{e^{z_k}(1 + \sum_{l \neq k} e^{z_l - z_k})} \approx \frac{e^{z_j}}{e^{z_k}} = e^{-(z_k - z_j)} = e^{-z_k - z_{j^*}}$$

This approximation holds because $\sum_{l \neq k} e^{z_l - z_k} = \sum_{l \neq k} o_l/o_k \approx \epsilon_s/(1-\epsilon_s) \approx \epsilon_s\mathbf{l}1$. Substituting this into the norm approximation (59):

$$\|\mathbf{J}_{\mathbf{W}}\|_F^2 \approx \sum_{j \neq k} (e^{-z_k - z_{j^*}})^2 \|\boldsymbol{w}_j - \boldsymbol{w}_k\|_2^2 = \sum_{j \neq k} e^{-2z_k - z_{j^*}} \|\boldsymbol{w}_j - \boldsymbol{w}_k\|_2^2 \tag{60}$$

The specific margin defined as $z_k - z_{j^*} = z_k - z_{j^*}$, where j^* is the top competitor (highest logit z_j among $j \neq k$), corresponds to the term with the largest $e^{-2z_k - z_{j^*}}$ (smallest $z_k - z_{j^*}$) in the sum, which often dominates the sum's value.

Behavior of δ_{TCB}. As the margin $z_k - z_{j^*}$ increases, the corresponding $z_k - z_{j^*}$ for the closest competitors also increases. This leads to an exponential decrease in the terms $e^{-2z_k - z_{j^*}}$ in (60). Consequently, $\|\mathbf{J_W}\|_F^2$ decreases strongly (exponentially) as $z_k - z_{j^*}$ increases. Since $\delta_{\mathrm{TCB}} = \epsilon / \|\mathbf{J_W}\|_F$, an increase in $z_k - z_{j^*}$ causes a decrease in the denominator, leading to an *increase* in δ_{TCB}. **Conclusion (Low $\mathcal{V}_{\mathrm{eff}}$):** In the low-flatness regime, TCB increases rapidly as the logit margin $z_k - z_{j^*}$ increases:

$$\delta_{\mathrm{TCB}} \propto \frac{1}{\sqrt{\sum_{j \neq k} e^{-2(z_k - z_j)} \|\boldsymbol{w}_j - \boldsymbol{w}_k\|_2^2}} \approx \text{Increases strongly with } z_k - z_{j^*} \quad (\text{for } \mathcal{V}_{\mathrm{eff}} \approx 1)$$

This derivation provides the theoretical underpinning for the strong positive correlation $r_{\delta, z_k - z_{j^*}} \approx 0.62$ observed empirically in the low-$\mathcal{V}_{\mathrm{eff}}$ data subset (Table 1).

Negligible Influence of Residual $\mathcal{V}_{\mathrm{eff}}$ Variation. In this regime, $\mathcal{V}_{\mathrm{eff}}$ is already close to 1. Small changes in the probability distribution (e.g., caused by changing $z_k - z_{j^*}$) lead to minuscule changes in $\mathcal{V}_{\mathrm{eff}}$. Specifically, $\mathcal{V}_{\mathrm{eff}} = 1/S_2 = 1/(o_k^2 + \sum_{j \neq k} o_j^2)$. As $o_k \approx 1$ and $o_j = O(\epsilon_s)$, we have $\mathcal{V}_{\mathrm{eff}} \approx 1/(1 - 2\epsilon_s + O(\epsilon_s^2))$. While changes in $z_k - z_{j^*}$ affect ϵ_s and thus cause small fluctuations in $\mathcal{V}_{\mathrm{eff}}$, these variations are vastly outweighed by the direct exponential impact of $z_k - z_{j^*}$ on $\|\mathbf{J_W}\|_F^2$ via (60). This explains why the correlation $r_{\delta, \mathcal{V}_{\mathrm{eff}}}$ drops to nearly zero ($\approx$ 0.08) in the low-$\mathcal{V}_{\mathrm{eff}}$ regime.