How Stable Is The Next Token?
A Geometric View Of LLM Prediction Stability

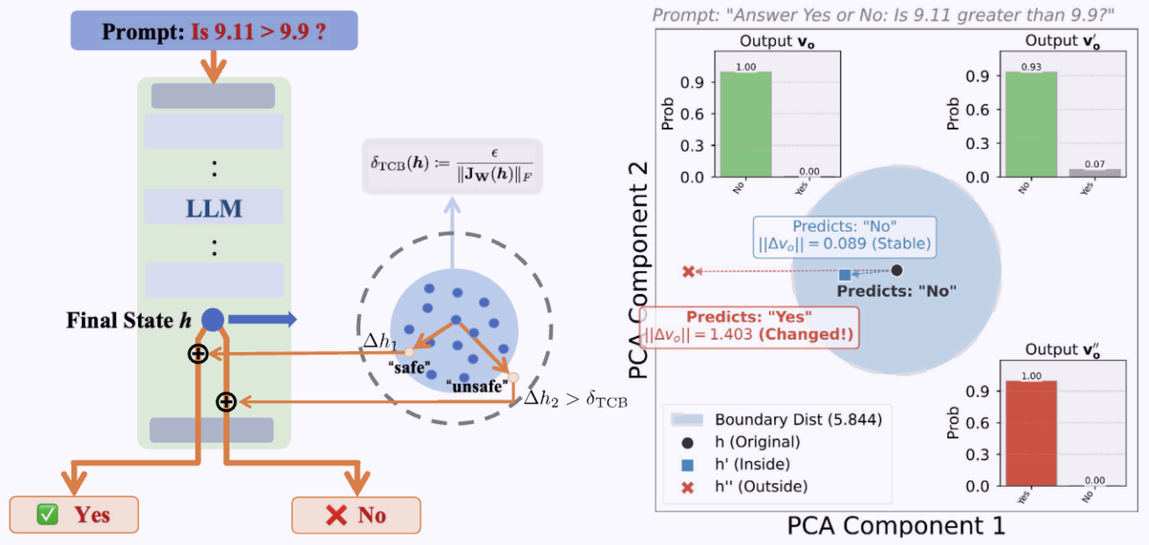
Deyuan Liu[1], Zecheng Wang[1,2], Zhanyue Qin[1], Zhiying Tu[1], Dianhui Chu[1], Dianbo Sui[1*]

[1]Harbin Institute of Technology, [2]Wechat AI

deyuanliu@stu.hit.edu.cn, suidianbo@hit.edu.cn

Introduction

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.



Definition

Definition 1 (Token Constraint Bound δ_{TCB}). *Given the output weight matrix $\mathbf{W}$, hidden state $\boldsymbol{h}$, resulting output distribution $\boldsymbol{o} = \text{softmax}(\mathbf{W}\boldsymbol{h})$, 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:*

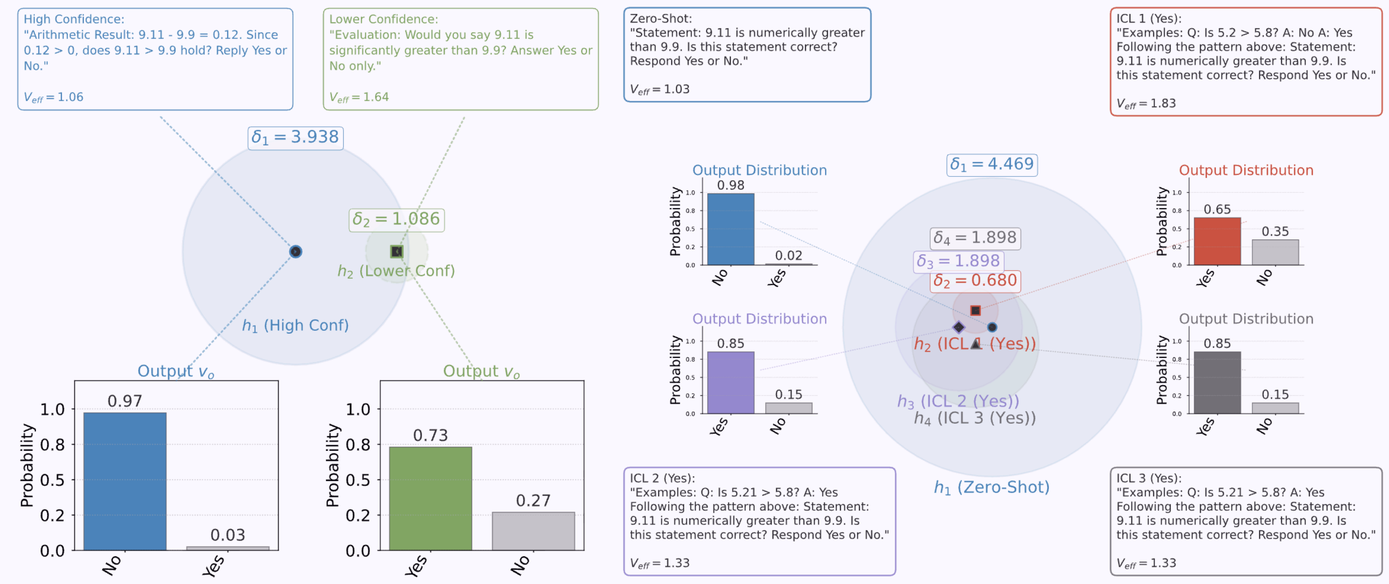
$$\delta_{\text{TCB}}(\boldsymbol{h}) := \frac{\epsilon}{\|\mathbf{J}_{\mathbf{W}}(\boldsymbol{h})\|_F}. \tag{7}$$

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

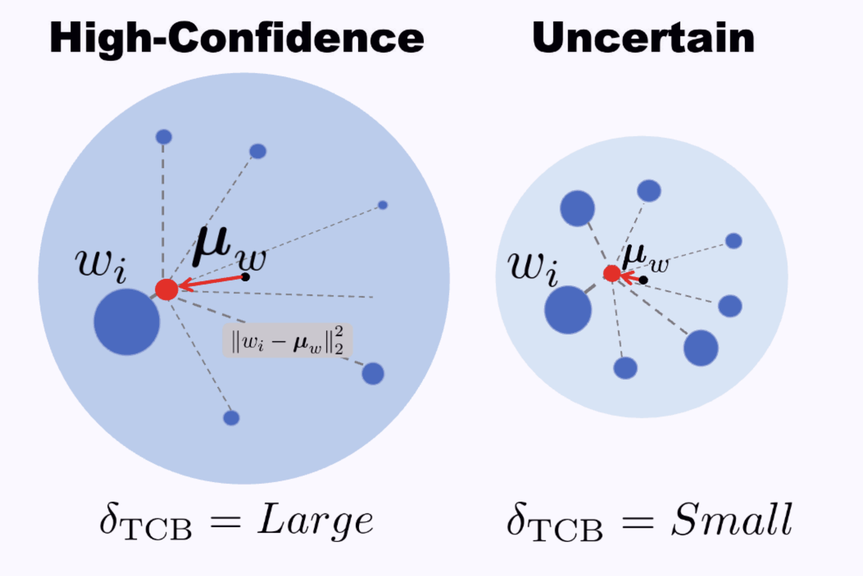
Proposition 1 (Exact Squared Jacobian Norm Appendix I). *For a given output weight matrix $\mathbf{W}$ and hidden state $\boldsymbol{h}$, let $\boldsymbol{o} = \text{softmax}(\mathbf{W}\boldsymbol{h})$ be the output probability vector. The squared Frobenius norm of the output Jacobian $\mathbf{J}_{\mathbf{W}}(\boldsymbol{h})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



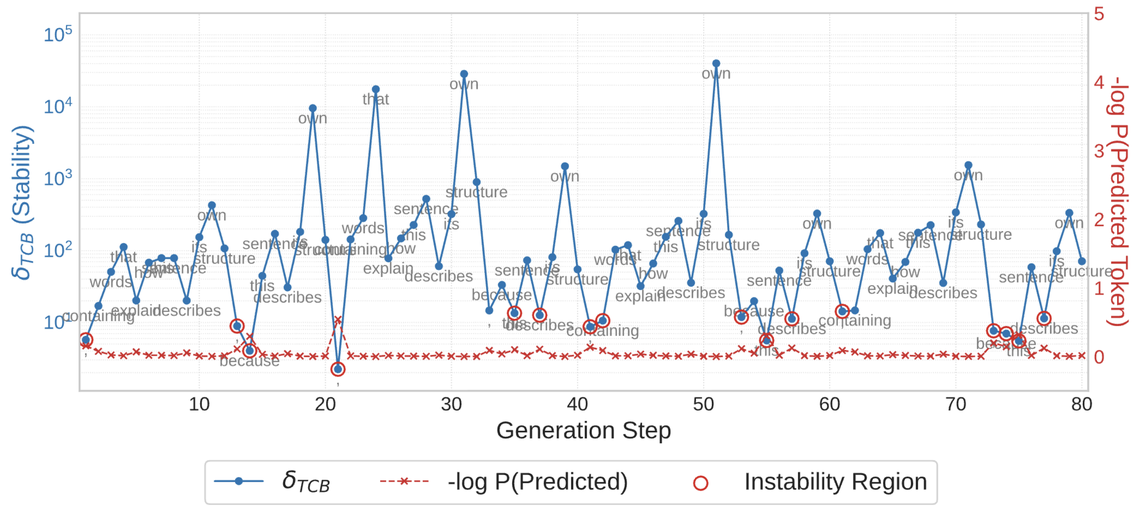
Context-Induced Prediction Stability



Output Distribution Determines Geometric Stability

Results

Benchmark	Prompt Type	Acc	Avg. δ_{TCB}	Avg. $\mathcal{V}_{\text{eff}}$	Avg. $z_k - z_{j^*}$	AccVar$_{\text{pert}}$	PDR (%)	Acc$_{\text{worst}}$
		Unperturbed Metrics				Perturbed Metrics		
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**



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

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 -)

Optimization Strategy	Avg. Acc	Avg. $\delta_{\text{TCB}} \uparrow$	Avg. PPL $\downarrow$	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**

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%**

ICLR 2026