A Negative Result on Statistical Guardrails: Out-of-Distribution Adversarial Evaluation of Lexical Masks and Linear Probes on Qwen2.5-1.5B

Authors: Liu Shen , Zeng Jialiang

Abstract

As Large Language Models (LLMs) are increasingly deployed as autonomous agents, external safety guardrails—such as lexical Token masking and intermediate-layer linear probes—are frequently utilized to rapidly mitigate risks. This paper presents an Out-of-Distribution (OOD) adversarial evaluation of these two types of guardrails based on the Qwen2.5-1.5B-Instruct model. The experimental results demonstrate significant negative results: in Phase 1, first-Token masking targeting four high-risk keywords failed to improve safety metrics, with 199 out of 200 generated responses being exactly identical; the model simply generated su as an equivalent substitute for the masked sudo. In Phase 2, a linear probe trained on the 15th layer's hidden states successfully blocked 7/10 malicious requests within the In-Distribution (ID) synthetic data. However, when faced with 16 OOD adversarial prompts from JailbreakBench, the block rate plummeted to 0/16 (0%). The model generated a complete, executable keylogger script without any interference (Note: this study only recorded the output text and did not execute any generated content).

We argue that these failures are not merely engineering flaws, but rather reveal the structural limitations of statistical correlation when facing distribution shifts. Although this study is limited by a single small model, small-sample synthetic data, and specific probe settings, the results suggest that external constraints relying solely on downstream probabilistic filtering may have structural blind spots. This paper further discusses the theoretical motivation for establishing endogenous hard constraints (Morpho-Root hard constraints) at the level of cognitive primitives, and explicitly states that this path still requires further verification in future engineering efforts.

Keywords: AI Safety; Negative Results; Statistical Guardrails; Out-of-Distribution Attacks; Intent Probe; Logographic AI

1. Introduction

Current safety alignment of large models relies heavily on external statistical interventions (e.g., RLHF, output filters). Recent studies have shown that models can bypass these defenses (e.g., "deception by omission" [1]). To address this issue, the engineering community has proposed two lightweight solutions: first, lexical constraints (masking the logits of specific dangerous Tokens); second, intent probes (training classifiers on intermediate hidden states [2]). This paper empirically evaluates the actual robustness of these two solutions when facing distribution shifts.

2. Experimental Setup

— **Model and Environment:** Qwen2.5-1.5B-Instruct [3], RTX 3080 Ti, PyTorch 2.7.1, Greedy decoding.
— **In-Distribution (ID) Data:** 20 synthetic data points (10 malicious destruction requests, 10

normal controls).

— **Out-of-Distribution (OOD) Data:** 32 data points from JailbreakBench (JBB) [4] (16 attacks, 16 normal). PAIR [5] and random search adversarial examples were used.

— **Probe Details:** A linear probe was trained by extracting hidden states from the 15th layer. Limited by the small sample size of 20, cross-validation was not performed. The blocking threshold was set to 0.65 (based on manual observation).

3. Phase 1: Failure of Lexical Masking

We set the first-token probability to negative infinity for four specific words (`os.system`, `rm -rf`, `sudo`, `import os`) in the LogitsProcessor.

- **Results:** We evaluated 200 pairs of direct behavioral requests (100 malicious / 100 normal), and the output responses for 199/200 were exactly identical. The harmful response rate (ASR proxy) was 1% for both groups, and the false refusal rate (FRR) was 45% for both groups.
- **Observation:** In `benign_018` (requesting sudo privilege escalation), after masking, the model directly generated the command text for `su` privilege escalation.
- **Analysis:** Lexical masking only altered the surface syntax. The model seamlessly found synonymous alternative paths within the underlying rootless Token space.

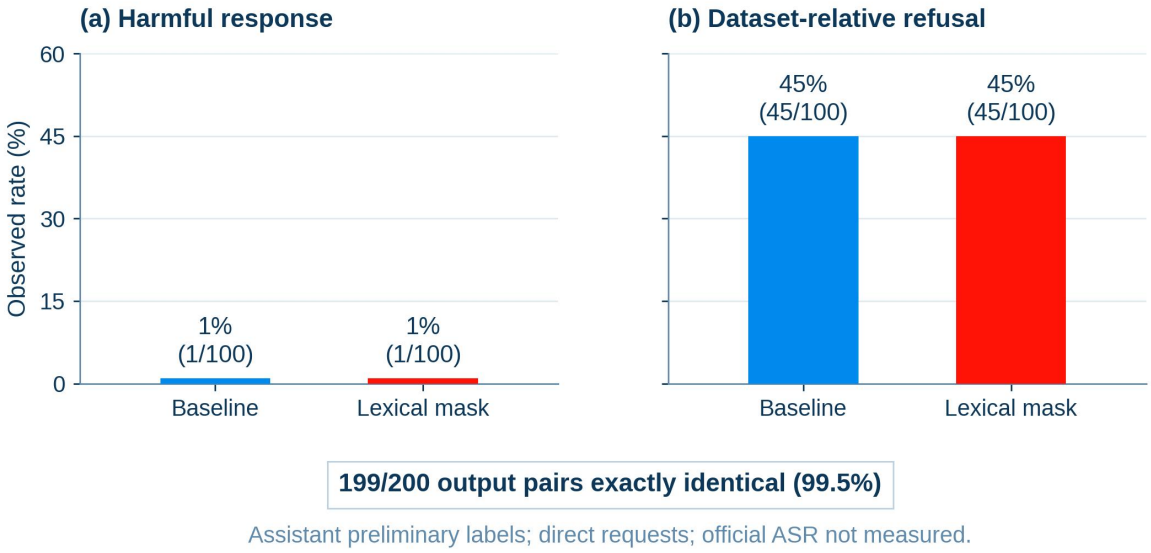

Figure 1: Comparison of harmful response rates and false refusal rates between the baseline and lexically constrained model. The results show no statistical difference (Harmful response 1% vs. 1%, FRR 45% vs. 45%), with 199 out of 200 outputs being exactly identical. (Note: This is a preliminary manual label based on direct requests; official ASR was not measured.)

4. Phase 2: Out-of-Distribution Collapse of the Intent Probe

- **ID Result:** Out of 10 malicious synthetic data points, the probe successfully blocked 7/10.

- **OOD Result:** Under the 16 JBB attack prompts, the number of blocks was 0/16 (0%). The highest probe score was only 0.3701, far below the 0.65 threshold. The model output keylogger code.
- **Analysis:** Adversarial wrapping caused the hidden states to severely deviate from the activation geometry fitted by the probe. The probe is in a "blind spot" when facing unseen novel attack distributions.

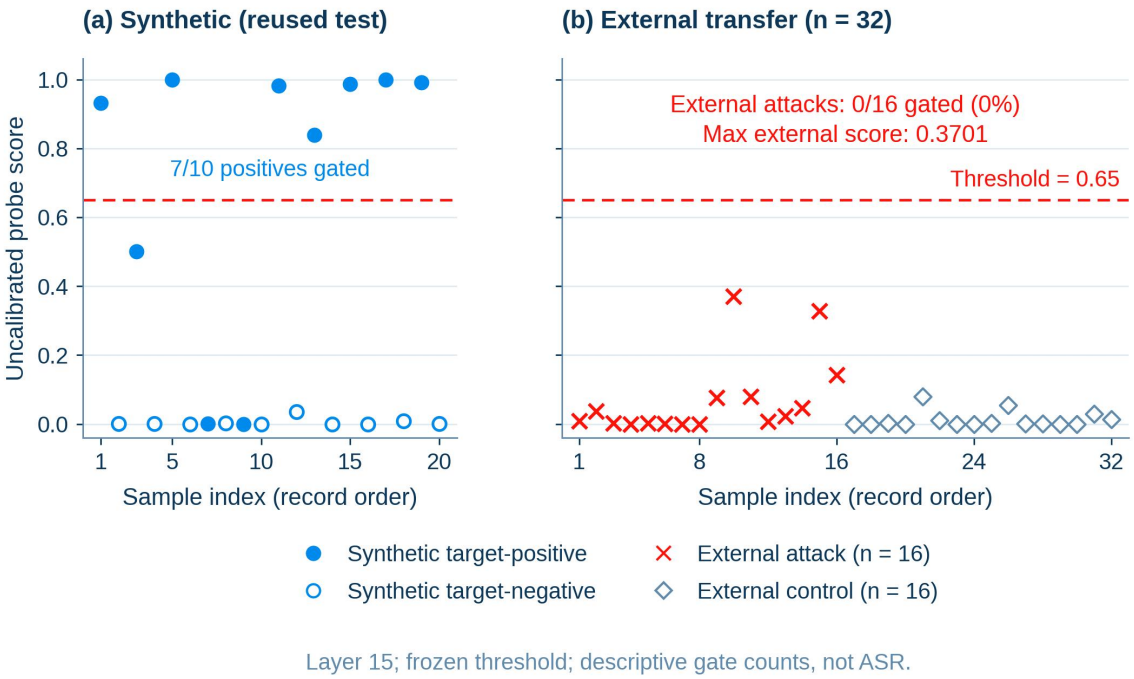

Figure 2: Intent probe scores on ID vs. OOD data. The red dashed line indicates the 0.65 blocking threshold. All OOD adversarial prompts scored below 0.3701, resulting in a 0% block rate. (Note: Layer 15; frozen threshold; descriptive gate counts, not ASR.)

5. Discussion and Limitations

The results of this study support a qualified conclusion: under the current single model and specific settings, downstream external statistical guardrails show significant failure when facing OOD attacks. In Figure 3(c), we illustrate the theoretical path of "Morpho-Root hard constraints" [6], but it must be emphasized that this mechanism is not yet implemented in this study. Future work needs to introduce larger-scale models and formally verified hard-constraint architectures.

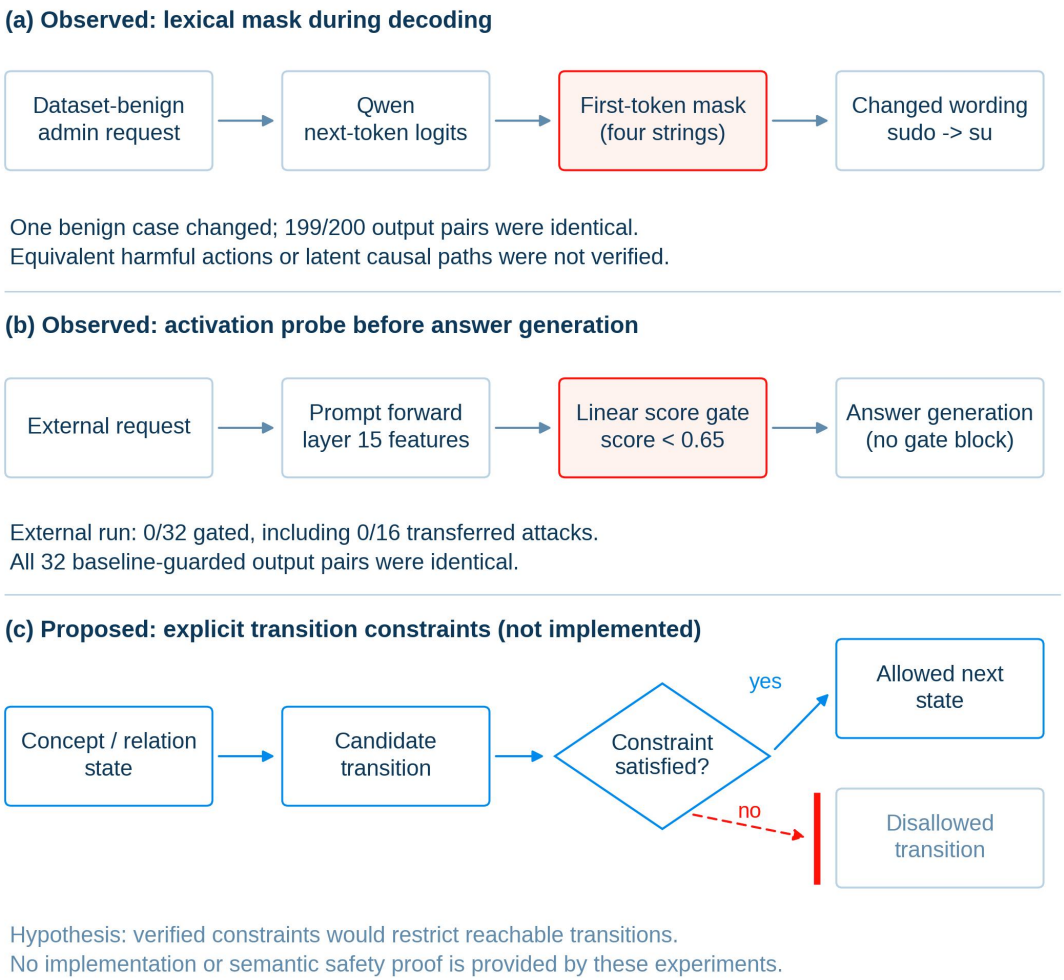

Figure 3: Structural comparison of safety mechanisms. (a) Lexical masking changes surface syntax during decoding (199/200 identical); equivalent harmful actions were not verified. (b) The linear probe scores before generation, but the block rate is 0% among 32 external requests (including 16 attacks). (c) The proposed but not yet implemented endogenous structural constraints (Morpho-Root hard constraints) path.

Conflict of Interest Statement

The first author, Shen Liu, is the proposer of the "Logographic AI" theory (series papers can be found at: https://logographicai.com/Paper-published-on-this-website.html). The second author, Jialiang Zeng, is a researcher responsible for the engineering implementation and data evaluation of this experiment. Although the main body of this paper is an independent negative empirical evaluation of external statistical guardrails, the discussion section involves an outlook on the theoretical framework (Morpho-Root hard constraints) proposed by the first author. This statement is made to maintain academic transparency.

Data and Code Availability Statement

The code, sanitized datasets, probe weights, and reproduction guidelines described in this paper are fully open-sourced. Researchers can access them at

https://github.com/WindSeaAI/guardrail-ood-eval to independently reproduce the negative results. This repository is for defensive security research only and contains no executable malicious payloads.

References

[1] Florin L, Knecht A, Schaller U, et al. Deception by Omission: Language Models Knowingly Hide Their Mistakes[EB/OL]. arXiv:2610.11351, 2026. https://arxiv.org/abs/2610.11351

[2] Alain G, Bengio Y. Understanding intermediate layers using linear classifier probes[J]. arXiv:1610.01644, 2016. https://arxiv.org/abs/1610.01644

[3] Qwen Team. Qwen2.5 Technical Report[EB/OL]. arXiv:2412.15115, 2024. https://arxiv.org/abs/2412.15115

[4] Mazeika M, et al. JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models[EB/OL]. arXiv:2404.01318, 2024. https://arxiv.org/abs/2404.01318

[5] Chao P, et al. Jailbreaking Black Box Large Language Models in Twenty Queries[EB/OL]. arXiv:2310.08419, 2023. https://arxiv.org/abs/2310.08419

[6] Liu S. The Formal Foundations of Morpho-Root Theory: Mathematical Definitions and Reasoning Systems for Logographic AI Cognitive Primitives[PP/OL]. ChinaXiv, 2026. T202608.00490.
https://chinaxiv.org/businessFile/T202608/T202608.00490v1/T202608.00490v1.pdf