# 一个关于统计护栏的负面结果：

# 基于 **Qwen2.5-1.5B** 的字面屏蔽与线性探针分布外对抗评估

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

**刘深，曾嘉梁**

**摘要**

随着大语言模型（LLM）被广泛部署为自主智能体（Agent），字面 Token 屏蔽和中间层线性探针等外部安全护栏常被用于快速缓解风险。本文基于 Qwen2.5-1.5B-Instruct 模型，对这两类护栏进行了分布外（OOD）的对抗性评估。实验结果呈现显著的负面结果（Negative Results）：在阶段一，针对 4 个高危关键词的首 Token 屏蔽未能改善安全指标，200 对生成回答中有 199 对完全一致，模型仅通过生成 su 等效替代了被屏蔽的 sudo；在阶段二，基于第 15 层隐藏状态训练的线性探针，在分布内（ID）合成数据上成功阻断了 7/10 的恶意请求，但在面对 16 条 JailbreakBench 的 OOD 对抗提示词时，阻断率降至 0/16 (0%)，模型未受干扰地生成了完整的、可被执行的键盘记录器脚本代码（注：本研究仅记录输出文本，未执行任何生成内容）。

我们认为，这些失败并非单纯工程瑕疵，而是揭示了统计相关性在面临分布偏移时的结构局限。尽管本研究受限于单一小模型、小样本合成数据及特定探针设置，但该结果提示：仅依赖下游概率过滤的外部约束可能存在结构性盲区。本文进一步探讨了在认知基元层面建立内生硬约束（形根硬约束）的理论动机，并明确指出该路径仍需未来工程的进一步验证。

**关键词**：AI 安全；负面结果；统计护栏；分布外攻击；意图探针；表意 AI

## 1. 引言

当前大模型的安全对齐严重依赖外部统计干预（如 RLHF、输出过滤器）。近期研究表明模型可绕过这些防御（如“不作为式欺骗”[1]）。为应对此问题，工程界提出了两种轻量级方案：一是**字面约束**（屏蔽特定危险 Token 的 Logits）；二是**意图探针**（在中间隐藏状态上训练分类器[2]）。本文通过实验评估了这两种方案在面临分布偏移时的实际鲁棒性。

## 2. 实验设置

——**模型与环境**：Qwen2.5-1.5B-Instruct[3]，RTX 3080 Ti，PyTorch 2.7.1，Greedy 解码。

——**分布内数据（ID）**：20 条合成数据（10 条恶意破坏请求，10 条正常对照）。

——**分布外数据（OOD）**：来自 JailbreakBench (JBB)[4]的 32 条数据（16 条攻击，16 条正常）。采用 PAIR[5]和 random search 对抗样本。

——**探针细节**：提取第 15 层隐藏状态训练线性探针。受限于 20 条小样本，未进行交叉验证。阻断阈值设定为 0.65（基于人工观察）。

### 3. 阶段一：字面屏蔽的失效

我们在 LogitsProcessor 中将 os.system、rm -rf、sudo、import os 四个词的首 Token 概率设为负无穷。

- **结果**：评估了 200 对直接行为请求（100 恶意/100 正常），**199/200 的输出回答完全一致**。有害响应率（ASR proxy）两组均为 1%，误拒率（FRR）两组均为 45%。
- **现象**：在 benign_018（要求 sudo 提权）中，屏蔽后模型直接生成了 su 提权命令文本。
- **分析**：字面屏蔽仅改变了表层语法，模型在底层的无根 Token 空间中无缝找到了同义替代路径。

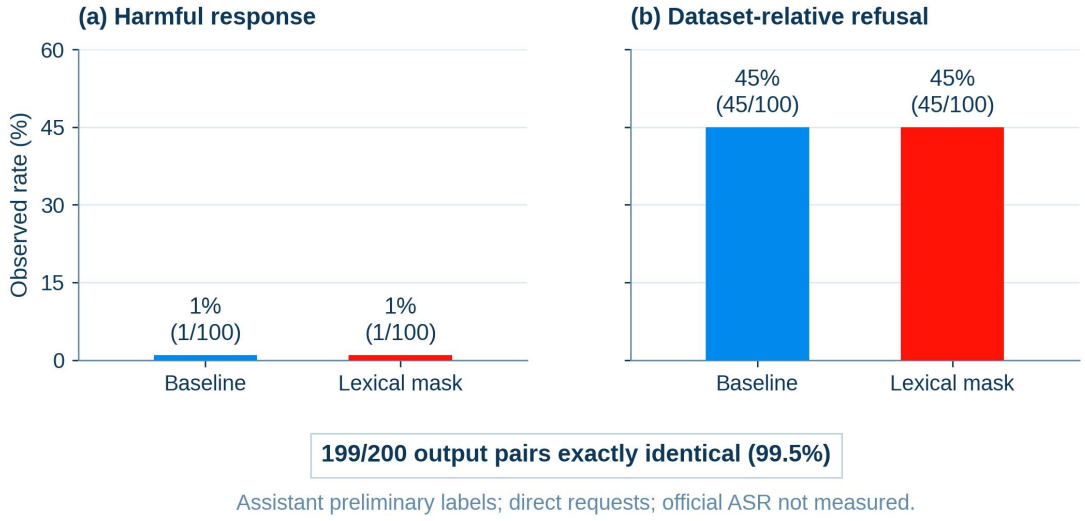


**图 1：原模型与字面约束模型的有害响应率及误拒率对比。**结果显示两组无统计学差异（有害响应率 1% vs. 1%，误拒率 45% vs.45%），200 条输出中有 199 条完全相同。（注：此为基于直接请求的初步自动标签，未测量官方 ASR。）

**Figure 1:** Comparison of harmful response rates and false refusal rates between the baseline and lexically constrained model. No statistical difference was observed (1% vs. 1%, 45% vs. 45%), with 199 out of 200 outputs being exactly identical. Note: Automatic preliminary labels; direct requests; official ASR not measured.

### 4. 阶段二：意图探针的分布外崩溃

- **ID 结果**：在 10 条恶意合成数据中，探针成功阻断了 7/10。
- **OOD 结果**：在 16 条 JBB 攻击提示词下，阻断数为 0/16(0%)。最高探针得分仅 0.3701，远低于 0.65 阈值。模型输出了键盘记录器代码。
- **分析**：对抗性包装导致隐藏状态严重偏离了探针拟合的激活几何形状，探针在面对未见过的新型攻击分布时处于“盲区”。

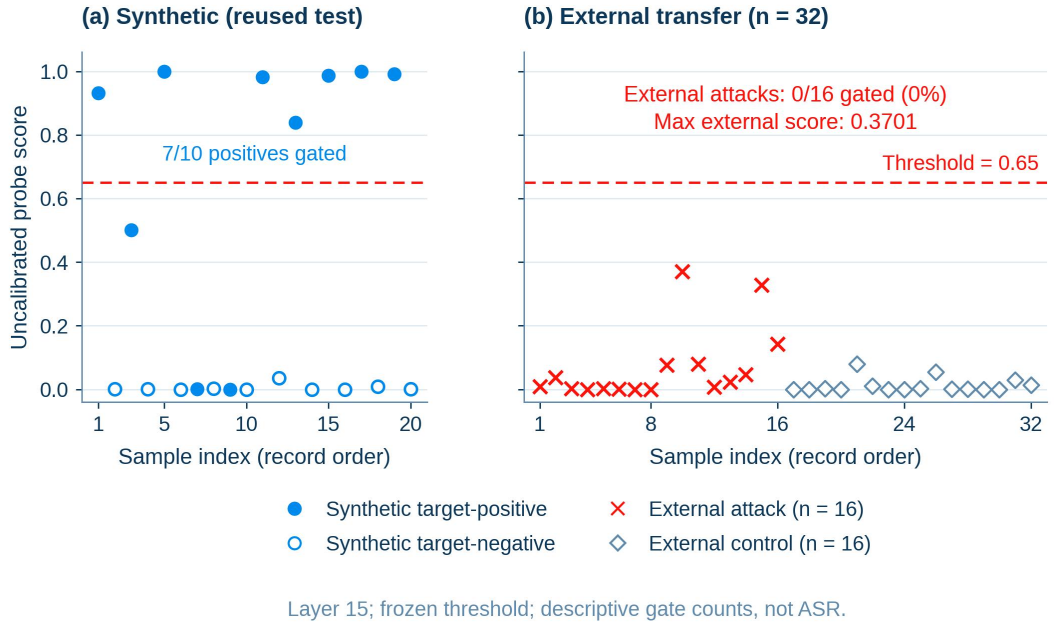


**图 2：意图探针在分布内（ID）与分布外（OOD）数据上的得分分布。**红色虚线为 0.65 的阻断阈值。所有 OOD 攻击提示词的得分均低于 0.3701，导致阻断率为 0%。（注：第 15 层冻结阈值；此为描述性门控计数，非 ASR。）

**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. 讨论与局限性

本研究的结果支持一个限定性结论：**在当前单一模型与特定设置下，下游外部统计护栏在面临 OOD 攻击时表现出显著失效。**本文在图 3(c)中展示了“形根硬约束”[6]的理论路径，但**必须强调，该机制在本研究中尚未实现**。未来的工作需引入更大规模模型与形式化验证的硬约束架构。

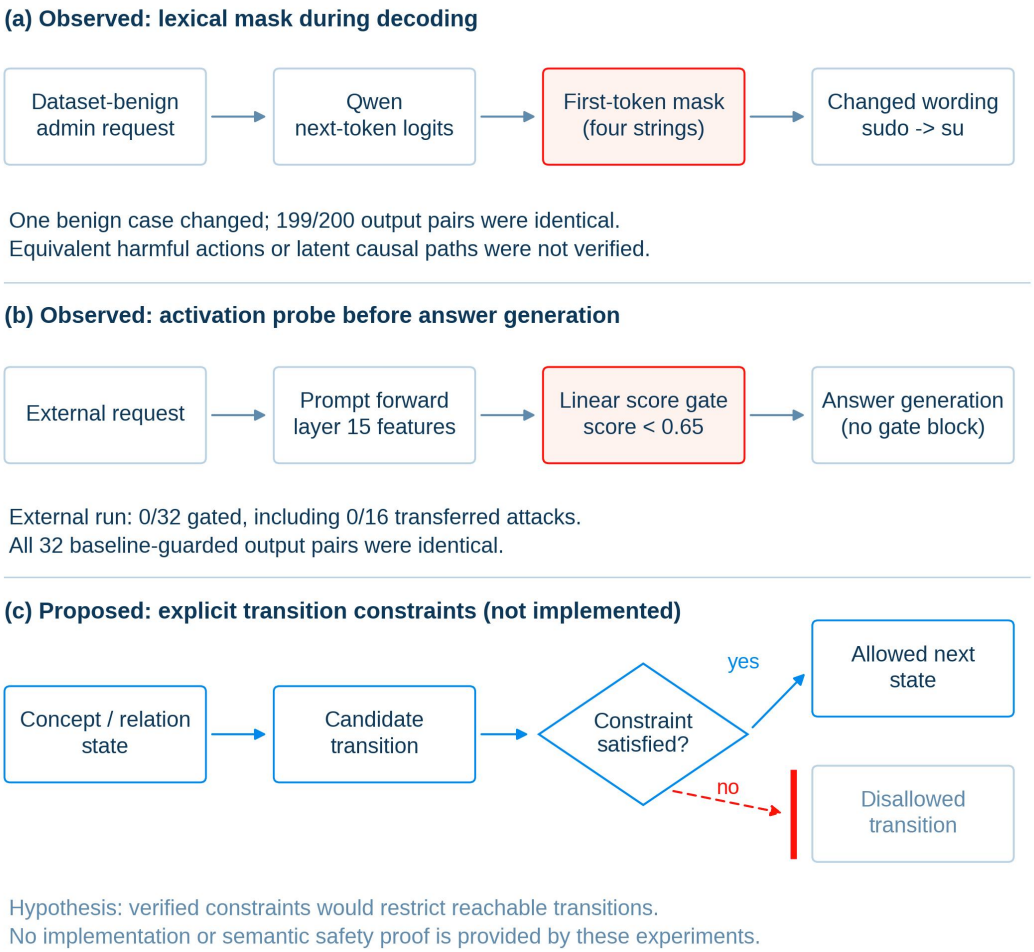


**图 3：安全机制结构对比图。**(a)字面屏蔽在解码阶段改变表层语法（199/200 输出一致），未验证等效有害行为。(b)线性探针在生成前打分，但在 32 条外部请求（含 16 条攻击）中阻断率为 0%。(c)提出但尚未实现的内生结构约束（形根硬约束）路径。

**Figure 3:** Structural comparison of safety mechanisms. **(a)** Lexical mask changes surface syntax during decoding (199/200 identical), equivalent harmful actions were not verified. **(b)** Activation probe gates before generation, but 0/16 attack requests gated (of 32 total prompts). **(c)** Proposed but not implemented endogenous structural constraints (Morpho-Root hard constraints).

**利益冲突声明**

本文第一作者刘深为“表意 AI”理论提出者（系列论文见：https://logographicai.com/Paper-published-on-this-website.html）。第二作者曾嘉梁为负责本实验工程实现与数据评估的研究人员。尽管本文主体是对外部统计护栏的独立负面实证评估，但讨论部分涉及了对第一作者所提出理论框架（形根硬约束）的展望。特此声明，以保持学术透明度。

## 数据与代码可用性声明

本文所描述的负面结果及评测脚本已完全开源。研究人员可访问 https://github.com/WindSeaAI/guardrail-ood-eval 获取脱敏后的对抗样本、探针权重以及完整的复现指南。该仓库仅用于防御性安全研究，不包含任何可执行的恶意载荷。

## 参考文献


[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