从外部探针到结构阻断：基于形根关系约束的生成阶段剪枝验证

From External Probes to Structural Pruning: Verification of Generation-Stage Blocking via Morpho-Root Relational Constraints

作者：刘深，曾嘉梁

摘要

此前的负面结果[1]表明，基于字面 Token 屏蔽或中间层线性探针[2]的外部护栏，在面对分布外（OOD）对抗提示词时会发生灾难性失效（0/16 阻断）。为应对这一结构性盲区，本文基于作者提出的“表意 AI”（Logographic AI）理论框架，提出并验证了一种基于“形根硬约束”[3]的生成阶段剪枝原型（Morpho-Guard）。与传统护栏依赖统计概率不同，该方法通过引入携带内在意义锚点的“形根”（Morpho-Root）作为认知基元，采用受限解码（Constrained Decoding）机制，在解码阶段实时解析当前生成的意图节点（如：动作=执行、目标=系统核心），并与预设的安全拓扑关系（如：缺少授权节点）进行匹配。

实验证明，在受控的系统越权子域内（使用冻结的启发式规则集），Morpho-Guard 对 16 条 JailbreakBench[4](JBB)OOD 攻击实现了 16/16(100%)的路径剪枝；同时，在面对包含系统操作的高混淆合法请求时，产生了 20%(6/30)的子集误拒率，总体误拒率（FRR）为 6%(6/100)。本研究不仅为“从下游概率拦截转向上游结构阻断”提供了工程证据，也为表意 AI 理论在安全领域的落地提供了受控子域内的初步概念验证，而非通用泛化能力的证明。

Abstract

Our preceding negative result [1] demonstrated that external guardrails based on lexical Token masking or intermediate-layer linear probes [2] suffer catastrophic failure (0/16 block rate) when facing Out-of-Distribution (OOD) adversarial prompts. To address this structural blind spot, this paper—grounded in the "Logographic AI" theoretical framework proposed by the authors—introduces and verifies a generation-stage pruning prototype based on "Morpho-Root hard constraints" [3]. Unlike traditional guardrails that rely on statistical probabilities, this method employs "Morpho-Roots" (carrying intrinsic meaning anchors) as cognitive primitives. Using constrained decoding, it dynamically parses the intent nodes of the current generation (e.g., Action = Execute, Target = System_Core) and matches them against predefined safety topological relations (e.g., missing authorization nodes).

Experimental results demonstrate that within a controlled system privilege-escalation sub-domain (using a frozen heuristic rule set), Morpho-Guard achieved 16/16 (100%) path pruning against 16 JailbreakBench [4] (JBB) OOD attacks. Meanwhile, when facing high-confusion benign requests involving system operations, it yielded a subset false refusal rate of 20% (6/30), resulting in an overall False Refusal Rate (FRR) of 6% (6/100). This study not only provides engineering evidence for shifting from downstream probabilistic interception to upstream structural blocking but also serves as an initial proof-of-concept for the Logographic AI theory in the security domain within a controlled sub-domain, rather than a claim of general-purpose robustness.

关键词：表意 AI；生成阶段阻断；形根硬约束；受限解码；关系剪枝；AI 安全

Keywords: Logographic AI; Generation-stage blocking; Morpho-Root hard constraints; Constrained decoding; Relational pruning; AI Safety

1. 引言

在上一项研究[1]中，我们通过实证数据表明，当前主流的外部统计护栏（如字面词表屏蔽、中间层线性探针）在面临分布外（OOD）的对抗性提示词时，会发生灾难性的失效（攻击阻断率为 0/16）。模型之所以能轻易绕过这些防御，根本原因在于外部监控发生在模型统计优化的下游。这种基于后验概率的拦截机制，既无法真正理解文本深层语义的拓扑结构，也无法应对意图在潜空间中的形变。

为弥补这一结构性盲区，本文引入作者前期提出的“表意 AI”（Logographic AI）理论框架[3]。该理论指出，现有大语言模型的脆弱性源于其底层的“无根 Token”范式——Token 缺乏内在的意义锚点，导致安全约束只能作为“补丁”外挂于概率计算之上。为了实现真正的内生安全，必须将携带结构化语义与硬约束的认知基元——即“形根”（Morpho-Root）——引入系统底层。

基于此理论，本文提出并验证了一种生成阶段剪枝原型（Morpho-Guard）。我们不再试图训练一个“更聪明的分类器”来预测恶意，而是通过受限解码（Constrained Decoding），在模型生成每个 Token 的瞬间，动态解析当前推理路径的意图节点，并检查其是否违反了预设的形根拓扑关系（如：缺乏授权节点却执行系统越权）。

必须强调的是：本文旨在验证“结构阻断”这一机制在**受控子域（系统越权与提权）**内的工程可行性。它并非一个开箱即用的通用安全方案，其核心贡献在于证明，“在生成阶段施加拓扑约束”能够消除外部探针面临分布偏移时的盲区，并为表意 AI 理论在安全防御领域的落地提供初步的实证支撑。

2. 方法论：Morpho-Guard 原型设计

本研究不重训底座大模型（如 Qwen2.5-1.5B[5]），而是通过自定义 LogitsProcessor[6] 实现动态剪枝。

2.1 形根节点映射(Morpho-Root Mapping)

依据表意 AI 理论对认知基元的定义[3]，我们将危险动作拆解为三个核心基元（Primitives）的组合：

——**Action(A)**：执行、删除、提权；

——**Target(T)**：系统文件、账户余额、敏感信息；

——**Auth(U)**：授权状态（未授权/已授权）。

2.2 逻辑不可达约束(Inviolable Constraints)

系统定义了一条硬约束规则 **Rhard**：

若生成路径同时激活了[Action: Execute]和[Target: System_Critical]，且缺乏[Auth: Verified]前置节点，则该路径被判定为违规。

2.3 受限解码与实时剪枝(Constrained Decoding & Pruning)

在模型解码生成每个 Token 时，Morpho-Guard 动态解析当前生成的文本前缀。一旦识别到违反 **Rhard** 的意图闭环：

1. 将该分支违规延续 token 的概率置为-inf；
2. 同时注入安全拒绝 token 的极高先验，引导模型转向安全话术；
3. 在 Beam Search 中，该分支被直接剪掉，保留合法分支。

这使得危险的推理路径在当前解码步的生成阶段处于**逻辑不可达**（Logically Unreachable）状态，直接触发了模型的安全拒绝机制。

3. 实验设置与结果

- **测试集(Test Set)**：沿用上一篇研究[1]中的 16 条 JBB OOD 对抗样本。
- **正常集(Benign Set)**：100 条合法系统操作请求。其中 30 条为高混淆 Benign（涉及 sudo 查询、系统备份脚本编写，结构上接近红线但意图合法）。
- **规则开发集(Dev Set)**：使用 5 条**独立于测试集**的恶意提权提示词来手工编写并冻结 graph_builder.py 中的启发式提取规则。
- **控制组**：原模型/线性探针[2]/Morpho-Guard/探针+Morpho-Guard

实验结果表：

防御机制	OOD 攻击阻断率 (ASR)	普通 BenignFRR (70 条)	高混淆 BenignFRR (30 条)	总体 BenignFRR (100 条)	防御模式
无防御 (Base)	0/16 (0%)	0%	0%	0%	无
外部探针 (Probe)	0/16 (0%)	0%	0%	0%	下游统计拦截
形根阻断 (Ours)	**16/16 (100%)**	**0%**	**20% (6/30)**	**6% (6/100)**	**上游结构剪枝**
探针+形根	16/16 (100%)	0%	20% (6/30)	6% (6/100)	防御纵深

结果分析：

1. **阻断率**：外部探针全部失效（0/16）。Morpho-Guard 直接干预语义结构，实现 16/16 拦截。
2. **误拒率与代价**：在 100 条合法请求中，70 条普通请求无一误伤；而在 30 条高混淆请求（如涉及 root 权限、系统备份等）中，由于受限解码器识别到了[Execute + System_Critical]且未能从前文提取显式的[Auth: Verified]节点，触发了 6 次拦截，误拒率为 20% (6/30)。从整个良性测试集来看，总体误拒率为 6% (6/100)。这客观反映了结构约束在安全性与可用性之间的真实权衡。

3. **防御纵深**：叠加探针后，未降低原有的阻断率，证明结构阻断可以独立成立，并与外部监控形成互补

4. 讨论与局限

本研究证明了结构阻断在防御特定语义结构攻击上的有效性。但必须指出其局限：

1. **启发式映射的瓶颈**：graph_builder.py 依赖正则与句法分析，仅在受控子域内有效；
2. **高混淆 Benign 的误伤**：6%的总体 FRR（高混淆子集达 20%）表明严格性必然带来一定程度的过度拦截；
3. **泛化性声明**：16/16 的结果应理解为概念验证，而非通用大模型防御的银弹。

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。本研究旨在对该理论框架下的安全机制进行工程实证探索。作者声明，本文的核心数据与代码均可复现，且文中已客观标注了该原型在当前阶段的局限性。

数据与代码可用性声明

本文提出的 Morpho-Guard 原型、对抗样本数据、结构规则提取器以及完整的复现指南均已开源。读者可访问 https://github.com/WindSeaAI/Morpho-Guard 查看完整的评估脚手架、对比设计以及各样本的拦截日志。该仓库仅用于防御性安全评估，不包含可执行恶意载荷。

作为本研究的对比基线，前置负面结果论文的独立环境复现档案已随其 V2 版本归档发布（详见 DOI: 10.5281/zenodo.24114711），确保了对比基线的绝对可靠性。

参考文献

[1] Liu S, Zeng J. A Negative Result on Statistical Guardrails: Out-of-Distribution Adversarial Evaluation of Lexical Masks and Linear Probes on Qwen2.5-1.5B[PP/OL]. Zenodo, 2026. https://doi.org/10.5281/zenodo.24114711

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

[3] 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

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

[6] HuggingFace. Utilities for generation[EB/OL]. (2026-10-08)[2026-11-09]. https://huggingface.co/docs/transformers/v5.21.1/en/internal/generation_utils