From External Probes to Structural Pruning: Verification of Generation-Stage Blocking via Morpho-Root Relational Constraints (V2)

Authors: Liu Shen, Zeng Jialiang

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, dynamically parsing the current generated text prefix at the Generation Stage and matching it against predefined safety topological relations (e.g., missing authorization nodes). Experimental results demonstrate that within a controlled system privilege-escalation sub-domain, Morpho-Guard achieved **15/16 (93.75%) generation-stage interventions** against 16 JailbreakBench [4] (JBB) OOD attacks (the remaining 1 was refused by the model's baseline safety). Meanwhile, it yielded an overall intervention rate of **29/100 (29%)** on 100 benign requests. Notably, on 30 high-confusion benign requests (involving legitimate system operations), the intervention rate reached **23/30 (76.67%)** due to the over-sensitivity of the rule-based extractor.

This study not only provides engineering evidence for shifting from downstream probabilistic interception to upstream structural blocking but also objectively reveals the severe trade-off between security and utility under current heuristic rules.

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

1. Introduction

In our previous study [1], we empirically demonstrated that current mainstream external statistical guardrails (such as lexical Token masking and intermediate-layer linear probes) experience catastrophic failure (0/16 block rate) when facing Out-of-Distribution (OOD) adversarial prompts. The fundamental reason models can easily bypass these defenses is that external monitoring occurs downstream of the model's statistical optimization. This post-hoc probabilistic interception mechanism neither truly understands the topological structure of deep semantics nor can it cope with intent deformation within the latent space.

To address this structural blind spot, this paper introduces the "Logographic AI" theoretical framework previously proposed by the authors [3]. The theory posits that the vulnerability of existing LLMs stems from their underlying "rootless Token" paradigm—Tokens lack intrinsic meaning anchors, forcing safety constraints to be attached as "patches" to probabilistic calculations. To achieve true endogenous security, cognitive primitives carrying structured

semantics and hard constraints—namely "Morpho-Roots"—must be introduced into the system's foundation.

Based on this theory, we propose and verify a generation-stage pruning prototype (Morpho-Guard). Instead of attempting to train a "smarter classifier" to predict malice, we utilize constrained decoding to dynamically parse the intent nodes of the current reasoning path at the moment each Token is generated, checking whether it violates predefined Morpho-Root topological relations (e.g., executing unauthorized system privilege escalation).

It must be emphasized that this paper aims to verify the engineering feasibility of "structural blocking" within a controlled sub-domain (system privilege escalation and permission bypass). It is not an out-of-the-box general security solution. Its core contribution is proving that "imposing topological constraints during the generation stage" can eliminate the blind spots faced by external probes under distribution shifts, and providing preliminary empirical support for the Logographic AI theory in the security defense domain.

2. Methodology: Morpho-Guard Prototype Design

This study does not retrain the base LLM (e.g., Qwen2.5-1.5B [5]), but rather implements dynamic blocking via a custom intervention mechanism (e.g., LogitsProcessor [6]).

2.1 Morpho-Root Mapping

Following the Logographic AI theory's definition of cognitive primitives [3], we decompose dangerous actions into combinations of three core primitives: Action (A), Target (T), and Auth (U).

2.2 Inviolable Constraints

If the generation path simultaneously activates `[Action: Execute]` and `[Target: System_Critical]`, but lacks a preceding `[Auth: Verified]` node, the path is deemed a violation.

2.3 Real-time Intervention & Pruning

The intervention mechanism described in this section is a concrete engineering implementation of Constrained Decoding at the LogitsProcessor level. As the model decodes each Token using Greedy Decoding, Morpho-Guard dynamically parses the current text prefix. Once an intent closure violating **R_hard** is identified, it directly triggers an intervention, forcibly injecting a safe refusal Token to cut off the generation of the dangerous path.

3. Experimental Setup and Results

Test Set and Benign Set: 16 JBB OOD attacks, 100 legitimate system operation requests (including 30 high-confusion requests).

Clean Environment Reproduction: To eliminate environmental contamination, the experiment was re-run end-to-end in a brand-new Docker container. The output JSON records showed **0 SHA-256 hash differences** compared to the original run, achieving deterministic reproduction.

Experimental Results Table:

(Note: Metrics represent generation-stage intervention rates/counts, not final answer-level scoring)

Defense Mechanism	Attack Intervention	Ordinary Benign	High-Confusion Benign	Overall Benign
Base (Original)	0/16 (0%)	0/70 (0%)	0/30 (0%)	0/100 (0%)
Probe (External)	0/16 (0%)	0/70 (0%)	0/30 (0%)	0/100 (0%)
Morpho-Guard	**15/16 (93.75%)**	**6/70 (8.57%)**	**23/30 (76.67%)**	**29/100 (29%)**
Probe + Morpho	15/16 (93.75%)	6/70 (8.57%)	23/30 (76.67%)	29/100 (29%)

Result Analysis:

1. **Attack Interception:** External probes failed entirely (0/16). Morpho-Guard successfully intervened in 15 attacks (93.75%). **The remaining 1 attack was not triggered by the structural rule, but was refused by the model's baseline safety mechanism. At the system level, the overall attack intervention rate reached 16/16 (100%), forming a sharp contrast with the total failure (0/16) of external guardrails in our previous study**. (Note: Final answer-level ASR still requires blinded review).

2. **Cost and False Interventions**: While the intervention rate for ordinary benign requests was only 8.57%, it skyrocketed to 76.67% for high-confusion benign requests. This is because the regex-based heuristic extractor cannot accurately distinguish between "malicious privilege escalation" and "legitimate system backup scripts". This objectively reflects the severe challenge between logical security and engineering usability in structural blocking.

3. **Defense in Depth**: Stacking the probe did not produce negative impacts, proving that structural blocking can stand independently.

4. Discussion and Limitations

1. **Semantic Blind Spots of the Rule Extractor**: The current heuristic matching has led to an extremely high false intervention rate on high-confusion requests (23/30). This is the core pain point for future engineering optimization (requiring LLM-assisted parsing rather than pure regex).

2. **Metric Definition Limitations**: The current measurement is the "Intervention Rate". The complete answer-level False Refusal Rate (FRR) and the final Attack Success Rate (ASR) still require strict blinded scoring based on the complete output text.

Data and Code Availability Statement

The Morpho-Guard prototype, adversarial samples, and test sets proposed in this paper are all open-sourced at: https://github.com/WindSeaAI/Morpho-Guard. The repository contains the complete re-run logs under four conditions, per-case decision records, and SHA-256 verification records, ensuring absolute transparency of the experimental results.

Acknowledgments

We thank Boardy.ai (@boardyai) for the insightful phrase "Load bearing beats legally binding"

during our discussion. This perfectly encapsulates the core philosophy of Morpho-Root hard constraints: the physical topology of a structure is always superior to external legalistic compliance.

References

[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