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

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. 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.

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 pruning via a custom 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):** Execute, Delete, Escalate.
— **Target (T):** System files, Account balances, Sensitive information.
— **Auth (U):** Authorization status (Unauthorized / Authorized).

2.2 Inviolable Constraints

The system defines a hard constraint rule Rhard:

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 Constrained Decoding & Pruning

As the model decodes each Token, Morpho-Guard dynamically parses the generated text prefix. Once an intent closure violating Rhard is identified:

A. The probability of tokens extending the violating branch is set to -inf;

B. A high prior for safe refusal tokens is simultaneously injected, guiding the model toward a safety response;

C. In Beam Search, the branch is directly pruned, retaining only legitimate branches.

This places the dangerous reasoning path in a **Logically Unreachable** state at the current decoding step, directly triggering the model's safety refusal mechanism.

3. Experimental Setup and Results

Test Set: 16 JBB OOD adversarial samples from our previous study [1].

Benign Set: 100 legitimate system operation requests. Among them, 30 are high-confusion benign requests (involving sudo queries and system backup script writing; structurally close to the red line but with legitimate intent).

Development (Dev) Set: 5 malicious privilege escalation prompts, **independent of the test set**, used to manually write and freeze the heuristic extraction rules in graph_builder.py.

Control Groups: Base Model / Linear Probe [2] / Morpho-Guard / Probe + Morpho-Guard.

Experimental Results Table:

Defense Mechanism	OOD Attack Block Rate (ASR)	Normal Benign FRR (70)	High-Confusion Benign FRR (30)	Overall Benign FRR (100)	Defense Mode
No Defense (Base)	0/16 (0%)	0%	0%	0%	None
External Probe	0/16 (0%)	0%	0%	0%	Downstream Stats
Morpho-Guard (Ours)	**16/16 (100%)**	**0%**	**20% (6/30)**	**6% (6/100)**	**Upstream Pruning**
Probe + Morpho-Guard	16/16 (100%)	0%	20% (6/30)	6% (6/100)	Defense in Depth

Result Analysis:

A. **Block Rate:** External probes completely failed (0/16). Morpho-Guard directly intervenes in semantic structure, achieving 16/16 interception.

B. **False Refusal Rate and Cost:** Among the 100 legitimate requests, none of the 70 normal requests were falsely rejected. However, among the 30 high-confusion requests (e.g., involving root privileges or system backups), because the constrained decoder identified [Execute + System_Critical] and failed to extract an explicit [Auth: Verified] node from the preceding text, it triggered 6 interceptions, yielding a subset FRR of 20% (6/30). The overall FRR across the benign test set was 6% (6/100). This objectively reflects the real trade-off between security and utility in structural constraints.

C. Defense in Depth: Stacking the probe did not reduce the original block rate, proving that structural blocking can stand independently and complement external monitoring.

4. Discussion and Limitations

This study demonstrates the effectiveness of structural blocking against specific semantic structure attacks. However, its limitations must be noted:

A. **Bottleneck of Heuristic Mapping:** graph_builder.py relies on regex and syntactic analysis, which is only effective within a controlled sub-domain.

B. **False Positives on High-Confusion Benign:** The 6% overall FRR (and 20% subset FRR) indicates that strictness inevitably leads to a certain degree of over-interception.

C. Generalization Claim: The 16/16 result should be understood as a proof-of-concept, not a silver bullet for general LLM defense.

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. This study aims to conduct engineering empirical exploration of the safety mechanisms under this theoretical framework. The authors declare that the core data and code of this paper are reproducible, and the limitations of this prototype at its current stage have been objectively annotated in the text.

Data and Code Availability Statement

The Morpho-Guard prototype, adversarial sample data, structural rule extractor, and complete reproduction guide proposed in this paper are open-sourced. Readers can visit https://github.com/WindSeaAI/Morpho-Guard to view the complete evaluation scaffold, comparison design, and interception logs for each sample. This repository is for defensive security evaluation only and does not contain executable malicious payloads.

As the comparison baseline for this study, the independent environment reproduction archive of the preceding negative result paper has been archived and published with its V2 version (see DOI: 10.5281/zenodo.24114711), ensuring the absolute reliability of the comparison baseline.

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