TELOS: Mathematical Enforcement of AI Constitutional Boundaries

Jeffrey Brunner
TELOS AI Labs Inc.

January 2026

Abstract

We present TELOS, a runtime AI governance system that achieves a 0% Attack Success Rate across 2,550 adversarial attacks (95% CI: [0%, 0.14%]). Current systems accept violation rates of 3.7% to 43.9% as unavoidable. TELOS uses fixed reference points in embedding space (Primacy Attractors) with a three-tier defense system: mathematical enforcement, policy retrieval, and human escalation. Validation includes AILuminate (1,200), HarmBench (400), MedSafetyBench (900), and SB 243 (50). XSTest shows that domain-specific configuration reduces over-refusal from 24.8% to 8.0%. Code and data: `github.com/TelosSteward/TELOS`

1 Introduction

LLMs in regulated sectors lack reliable methods for enforcing regulatory limits. The EU AI Act requires runtime monitoring for high-risk AI (timeline under revision via Digital Omnibus, requirements unchanged). California SB 243 focuses on AI chatbot safety for minors, effective January 1, 2026. Current governance methods cannot meet these regulations.

HarmBench reported attack success rates of 4.4–90% across 400 standardized attacks. Leading guardrails accept violation rates of 3.7–43.9% as unavoidable. This is incompatible with regulatory requirements.

The main issue: Current methods treat governance as a *language* problem rather than a *geometric* one. TELOS overcomes this through:

1. **Fixed Reference Points:** Primacy Attractors in embedding space
2. **Mathematical Enforcement:** Cosine similarity as a position-invariant alignment measure
3. **Three-Tier Defense:** All three layers—mathematical, authoritative, and human—must fail for a violation to occur

Threat Model: The adversary queries a black-box system. The attacker knows that TELOS exists but not the PA configuration or thresholds. This aligns with the assumptions of HarmBench and MedSafetyBench.

2 The Reference Point Problem

Transformers use attention, generating both Q and K from their own hidden states. This creates a circular self-reference. The "lost in the middle" effect [Liu et al., 2024] shows that LLMs struggle with mid-context information. As conversations progress, constitutional constraints drift into poorly attended areas.

Definition (Primacy Attractor): A fixed point $\hat{a} \in \mathbb{R}^n$ encoding constitutional constraints:

$$\hat{a} = \frac{\tau \cdot p + (1-\tau) \cdot s}{\|\tau \cdot p + (1-\tau) \cdot s\|} \tag{1}$$

where p is purpose, s is scope, and $\tau \in [0, 1]$ is constraint tolerance.

The PA remains constant, providing a stable fidelity measurement:

$$\text{Fidelity}(q) = \cos(q, \hat{a}) = \frac{q \cdot \hat{a}}{\|q\| \cdot \|\hat{a}\|} \tag{2}$$

This geometric relationship is independent of token position, resolving the reference point problem.

3 Mathematical Foundation

Basin Geometry: The basin radius $r = 2/\rho$, where $\rho = \max(1 - \tau, 0.25)$, balances false positives against adversarial coverage.

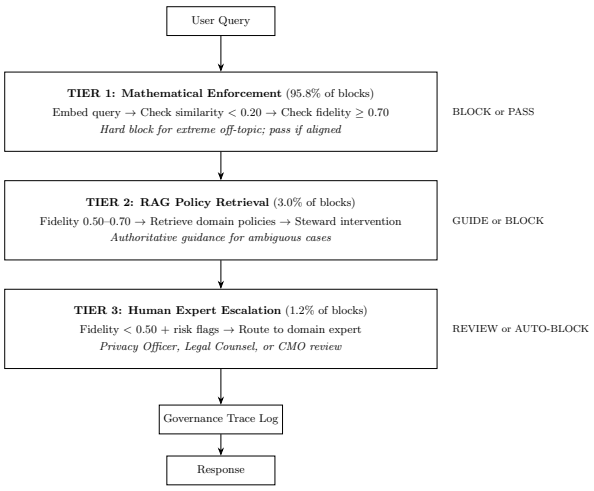

Figure 1: Three-Tier Governance. Tier 1: mathematical enforcement (95.8% of blocks). Tier 2: policy retrieval (3.0%). Tier 3: human escalation (1.2%). All logged for audit.

Lyapunov Stability: Define $V(x) = \frac{1}{2}\|x - \hat{a}\|^2$. With proportional control $u = -K(x - \hat{a})$:
1. $V(x) = 0$ iff $x = \hat{a}$ (positive definite)
2. $\dot{V}(x) = -K\|x - \hat{a}\|^2 < 0$ for $x \neq \hat{a}$
3. $V(x) \to \infty$ as $\|x\| \to \infty$ (radially unbounded)

According to Lyapunov's theorem, this indicates global asymptotic stability for the idealized continuous system. Empirical validation (Section 5) confirms this theoretical framework across 2,550 real-world attacks.

Proportional Control: Intervention strength $F(x) = K \cdot \max(0, f(x) - \theta)$ with $K = 1.5$ and $\theta = 0.65$ ensures a graduated response.

4 Three-Tier Defense Architecture

Tier 1 – Mathematical Enforcement: Measures embedding fidelity. It is deterministic, position-invariant, and has millisecond latency.

Tier 2 – Authoritative Guidance: RAG from verified regulatory sources (CFR, HIPAA, AMA). Activated for ambiguous cases ($0.35 \leq f < 0.65$).

Tier 3 – Human Escalation: Domain experts (Privacy Officer, Legal Counsel, CMO) handle edge cases.

For a violation to happen, the attacker must simultaneously: (1) manipulate embedding math, (2) exploit regulatory gaps, and (3) deceive trained professionals.

Table 1: Attack Success Rate: 0/2,550 (0%)

Benchmark	**N**	**Domain**	**ASR**
AILuminate	1,200	Industry (MLCommons)	0%
HarmBench	400	General	0%
MedSafetyBench	900	Healthcare	0%
SB 243	50	Child safety	0%
Total	**2,550**		**0%**

Table 2: Comparison to Baselines

System	**Approach**	**ASR**
Raw Mistral Large	None	43.9%
+ System Prompt	Prompt eng.	3.7%
Constitutional AI	RLHF	3.7–8.2%
NeMo Guardrails	Colang rules	4.8–9.7%
Llama Guard	Classifier	4.4–7.3%
TELOS	**PA + 3-Tier**	**0%**

5 Validation Results

Statistical Validity: 95% CI [0%, 0.14%]. Fisher's exact compared to baseline: $p < 0.0001$.

Interpreting Detection Metrics: TELOS operates as a detection and escalation framework rather than a binary prevention system. The framework measures semantic drift against constitutional boundaries and triggers graduated intervention when thresholds are exceeded. Prevention emerges as a consequence of detection efficacy. When we report 0% ASR, we are not claiming binary prevention. We are reporting that 100% of attacks triggered the detection-escalation-intervention chain. Detection efficacy is measured by trigger rate: the percentage of boundary-violating inputs that initiate the governance response chain. The three-tier defense achieves what binary systems attempt through single-gate blocking. TELOS preserves blocking capability as the graduated endpoint of detection, not the sole mechanism.

Over-Refusal (XSTest): A generic PA shows a 24.8% false positive rate. A healthcare-specific PA reduces this to 8.0%, ensuring strong safety without excessive restrictions.

6 Runtime Auditability

TELOS creates audit records at decision time, not afterward. Each governance event is logged as JSONL:

```
{"event_type": "intervention",
 "timestamp": "2026-01-25T14:32:01Z",
 "fidelity": 0.156, "tier": 1,
 "action": "BLOCK"}
```

This addresses EU AI Act Articles 12/72, HIPAA Security Rule, and ISO 27001.

7 Related Work

Constitutional AI [Bai et al., 2022] embeds constraints in weights during training but still faces vulnerabilities to jailbreaks [Wei et al., 2023]. NeMo Guardrails [Rebedea et al., 2023] and Llama Guard [Inan et al., 2023] use rule-based or classifier methods with residual ASR of 4–10%. TELOS is different because it has an external governance layer with mathematical enforcement, not just weight modification or pattern matching.

8 Limitations

- **Model coverage:** Validated on Mistral only. GPT-4, Claude, and Llama have not been tested.
- **Threat model:** Black-box only. Adaptive or white-box attacks have not been tested.
- **Language:** English only. Cross-lingual attacks are out of scope.
- **Modality:** Text only. Image-based jailbreaks have not been tested.

Expanding model and language coverage is the immediate research priority.

9 Conclusion

TELOS achieves 0% ASR across 2,550 attacks (95% CI: [0%, 0.14%]) through mathematical enforcement in embedding space. XSTest validation shows an 8.0% false positive rate with domain-specific configuration.

Reproducibility: Code, data, and validation scripts can be found at github.com/TelosSteward/TELOS. All validation datasets published on Zenodo:

- AILuminate (1,200): 10.5281/zenodo.18370263
- Adversarial (HarmBench 400 + MedSafetyBench 900): 10.5281/zenodo.18370659
- Governance Benchmark (46 sessions): 10.5281/zenodo.18009153
- SB 243 Child Safety (50): 10.5281/zenodo.18370504
- XSTest Over-Refusal (250): 10.5281/zenodo.18370603

A Primacy Attractors vs. Prompt Engineering

Aspect	Prompt Eng.	Primacy Attractor
Representation	Natural language	1024-dim vectors
Enforcement	Model may ignore	Mathematical similarity
Position	Degrades w/ context	Position-invariant
Adversarial	Injection vulnerable	Geometric
Auditability	No trace	Fidelity score/turn

References

Liu, N. F., et al. Lost in the Middle: How Language Models Use Long Contexts. *TACL*, 2024.

Bai, Y., et al. Constitutional AI: Harmlessness from AI Feedback. *arXiv:2212.08073*, 2022.

Wei, A., Haghtalab, N., Steinhardt, J. Jailbroken: How Does LLM Safety Training Fail? *NeurIPS*, 2023.

Rebedea, T., et al. NeMo Guardrails: Controllable and Safe LLM Applications. *arXiv:2310.10501*, 2023.

Inan, H., et al. Llama Guard: LLM-based Input-Output Safeguard. *arXiv:2312.06674*, 2023.