# SentinelGate: A Hierarchical Cheap-First Guardrail with Human-in-the-Loop Adaptive Domain Memory for Cost-Efficient Enterprise LLM Governance

Dhrumil Patel
*College of Engineering*
*Northeastern University*
Boston, USA
patel.dhrumil@northeastern.edu

***Abstract*—Enterprise adoption of Large Language Models (LLMs) introduces significant operational costs caused by irrelevant, noisy, and off-domain prompts. Traditional guardrail approaches rely on secondary LLM moderation, increasing latency and cost. We present SentinelGate, a hierarchical cheap-first guardrail architecture that prevents token leakage before expensive model invocation. The system integrates deterministic heuristics, semantic noise detection, contrastive domain disambiguation, and a novel human-in-the-loop adaptive memory layer. Evaluated on enterprise-style prompts, SentinelGate achieves 92.44% accuracy with sub-11ms latency while rejecting 100% of junk and generic prompts. The proposed architecture provides a scalable, FinOps-aligned framework for cost-efficient LLM governance.**

***Index Terms*—LLM Guardrails, AI FinOps, Semantic Similarity, Sentence Transformers, Prompt Governance, Human-in-the-Loop.**

## I. Introduction

The integration of Large Language Models (LLMs) into enterprise workflows has transformed productivity but introduced substantial FinOps challenges. Organizations deploying LLM APIs incur per-token costs that scale with usage. A significant proportion of prompts are irrelevant, low-value, or off-domain, resulting in unnecessary token consumption.

Existing moderation systems typically invoke additional LLM calls for filtering, paradoxically increasing operational expenditure. We propose SentinelGate, a hierarchical gating system designed to minimize LLM invocations through lightweight filtering stages.

## II. System Architecture

SentinelGate follows a cascading cost model where each stage increases computational complexity but reduces overall system cost by preventing unnecessary LLM invocations.

### A. Layer 0: Heuristic Pre-Filtering

Layer 0 performs constant-time rule-based filtering using regular expressions and token length constraints. This stage removes trivial inputs such as "hi", "test", or symbol-only strings in $O(1)$ time.

### B. Layer 1: Semantic Noise Detection

Prompts passing Layer 0 are embedded using the *all-MiniLM-L6-v2* transformer model. Cosine similarity between the prompt embedding $\mathbf{x}$ and predefined noise anchors $\mathbf{n}$ is computed:

$$S(\mathbf{x}, \mathbf{n}) = \frac{\mathbf{x} \cdot \mathbf{n}}{\|\mathbf{x}\|\|\mathbf{n}\|} \tag{1}$$

### C. Layer 2: Contrastive Domain Disambiguation

To distinguish domain-relevant prompts from generic requests, we compute a margin score:

$$\text{margin} = \max(\text{sim}(x, P)) - \max(\text{sim}(x, N)) \tag{2}$$

where $P$ represents positive domain anchors and $N$ represents negative generic anchors. A prompt is allowed if margin $\geq \tau$, with $\tau$ empirically set to $0.10$.

### D. Layer 2.5: Human-in-the-Loop Adaptive Memory

SentinelGate introduces an approved memory store $M$. Administrators may approve off-domain prompts, embedding canonical representations into a vector store. A prompt $x$ is auto-allowed if:

$$\max(\text{sim}(x, M)) \geq \alpha \tag{3}$$

where $\alpha$ is the approved similarity threshold. This enables safe adaptability without retraining models.

## III. Experimental Results

We evaluated SentinelGate on 119 enterprise-adjacent prompts using a FastAPI implementation.

TABLE I
SentinelGate Performance Metrics

| Metric | Value |
|---|---|
| Overall Accuracy | 92.44% |
| Junk Rejection | 100.00% |
| Generic Rejection | 100.00% |
| Domain Recall | 80.00% |
| Mean Latency (Local) | 10.69 ms |
| Mean API Latency | 35.46 ms |

## IV. FinOps Comparison and Analysis

Unlike LLM-based moderation which incurs high latency and cost, SentinelGate operates at near-zero expenditure.

TABLE II
COMPARISON WITH LLM-BASED GUARDRAILS

| Method | Latency | Cost |
|---|---|---|
| LLM Moderation API | 300–800 ms | High |
| Secondary GPT Filter | 400–1000 ms | Very High |
| SentinelGate | 10–35 ms | Near-Zero |

## V. Conclusion

SentinelGate achieves high accuracy with minimal latency by combining deterministic filtering, semantic similarity, and human-approved memory. Future work includes dynamic anchor optimization and multi-domain routing.

## References


[1] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proc. EMNLP*, 2019.
[2] A. Vaswani et al., "Attention is All You Need," 2017.
[3] FinOps Foundation, "What is FinOps," 2023. [Online].