Sentra: A Kubernetes Sidecar-Based Multi-Layer Security Gateway for Protecting LLM APIs[*]

Jaeyoung Lee, Chanuk Park, and Jaehyun Nam

Dankook University, Yongin, Gyeonggi-do, 16890, Republic of Korea
{leeja042499, cupark, namjh}@dankook.ac.kr

Abstract

This paper introduces *Sentra*, a Kubernetes sidecar-based security gateway designed to protect LLM APIs from jailbreak, model extraction, and DoS(Denial-of-Service) attacks. Unlike approaches requiring retraining or code modification, *Sentra* acts as a transparent reverse proxy implementing three defense layers for input filtering, extraction prevention, and availability protection. When *Sentra* was applied in a normal service environment, the average latency increased by about 12%, yet it effectively blocked service attacks, model extraction, and jailbreak attempts, enhancing overall security.

1 Introduction

Large Language Models (LLMs) expose new attack surfaces when served via public APIs [1]. Adversaries can exploit jailbreak prompts to bypass safety rules, issue repetitive queries to extract model knowledge, or flood requests to exhaust resources. Existing defenses such as fine-tuning, guardrail integration, or log-based monitoring suffer from high cost, poor scalability, and delayed response. To address these issues, we present *Sentra*, a Kubernetes sidecar-based security gateway that transparently protects LLM APIs at the infrastructure level. Operating independently from model logic, *Sentra* intercepts all API traffic and applies lightweight, multi-layer defenses without modifying the Inference container. This approach enables model-agnostic, real-time protection with minimal latency overhead.

2 Our Approach

Figure 1 shows the overall workflow of *Sentra*, which processes all API requests from external users before they reach the Inference container. When a user sends a request, the sidecar

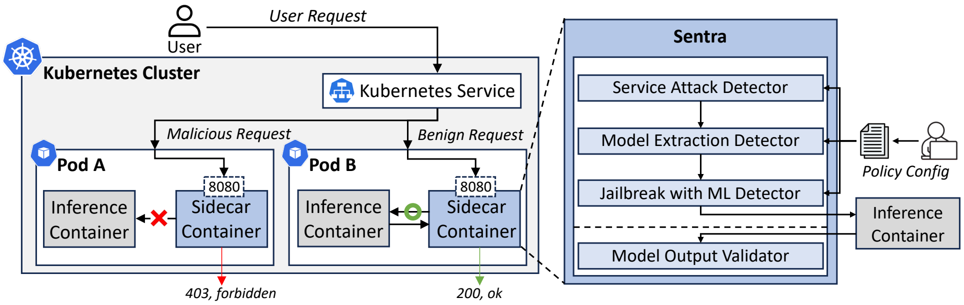

Figure 1: The Overall Architecture and Workflow of *Sentra* System

[*] Proceedings of the 9th International Conference on Mobile Internet Security (MobiSec'25), Article No. P-60, December 16-18, 2025, Sapporo, Japan. © The copyright of this paper remains with the author(s).

intercepts it and sequentially passes it through three detection modules, namely the Service Attack Detector, the Model Extraction Detector, and the Jailbreak Detector, each providing a dedicated protection layer. Only requests that pass all three stages are forwarded to the LLM, while any suspicious request is immediately and automatically blocked.

The Service Attack Detector and Model Extraction Detector jointly protect the model's availability and intellectual property. The Service Attack Detector limits request rates to prevent DoS(Denial-of-Service) attacks, with all policies configurable by the administrator. The Model Extraction Detector compares each prompt with recent ones from the same client and detects repeated similarities with minimal length variation, which indicates extraction attempts.[2] Thresholds and limits are configurable for different deployment environments.

The implemented Jailbreak Detector uses a machine learning–based classification model to detect prompt-injection and jailbreak attempts in user inputs. The model analyzes semantic and syntactic patterns to identify adversarial intent and compares text embeddings against a trained decision boundary to determine whether the input is benign or malicious. Detected jailbreak or injection attempts are immediately blocked, ensuring safe responses.

Figure 2 presents the performance evaluation of *Sentra* in a Kubernetes cluster with two NVIDIA RTX 4090 GPUs, measuring cpu usage, latency, throughput and execution time under 1, 2, and 4 concurrent requests. Results show that latency rose by about 12%, throughput dropped by roughly 8.9%, and cpu usage increased moderately, demonstrating that *Sentra* maintains stable performance while ensuring effective protection even under parallel workloads.

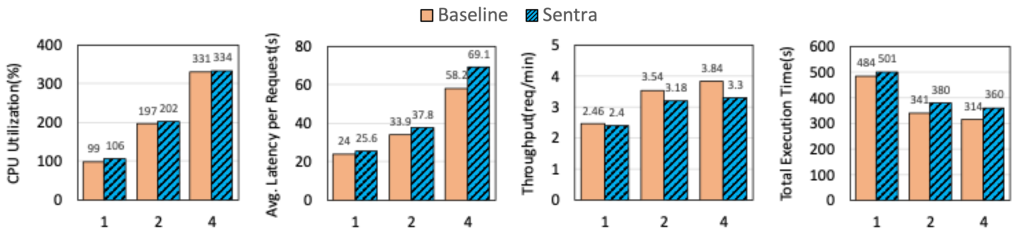

Figure 2: Performance measurements of *Sentra* under 1, 2, and 4 concurrent requests.

3 Conclusion

This paper presents *Sentra*, a Kubernetes sidecar–based multi-layer security gateway designed to protect LLM APIs. The system integrates detectors for service attacks, model extraction, and jailbreak attempts, providing transparent and model-independent protection. Evaluation results show that *Sentra* effectively mitigates diverse attack vectors while adding only 12% latency overhead, thereby enhancing the overall security of LLM-based services. Future work will extend detection to multimodal models and develop adaptive defense mechanisms to counter evolving threats.

Acknowledgement: This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT). (No.RS-2024-00398379, Development of High Available and High Performance 6G Cross Cloud Infrastructure Technology)

References

[1] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing Machine Learning Models via Prediction APIs. In *Proceedings of the USENIX security symposium*, pages 601–618, 2016.

[2] Z. Li, Y. Chen, W. Zhao, et al. Model Extraction Attacks Revisited. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024.