Carlos Denner dos Santos
Videns, propelled by Cofomo
Montreal, Quebec, Canada
carlos.denner@videns.ai

November 24, 2025

Editorial Board
Communications of the ACM
Association for Computing Machinery
New York, NY, USA

Dear Editor,

I am writing to submit our manuscript titled **"Prompt Injection Demystified: Building an LLM Firewall for Production LLM Systems"** for consideration as a Practice article in Communications of the ACM.

Prompt injection has emerged as OWASP's number one security risk for LLM applications, with recent incidents affecting major AI coding assistants (CVE-2025-54135, CVE-2025-54136) and enterprise copilots demonstrating the urgent need for practical defenses. Despite widespread industry concern—evidenced by 18 recent patent filings from OpenAI, Microsoft, Google, and Meta—practitioners currently lack rigorously evaluated, deployable solutions with quantified trade-offs.

This article addresses that gap by presenting a production-ready LLM firewall architecture evaluated across eight experimental phases with 925 total test cases. The firewall combines Unicode normalization, signature-based detection, and semantic screening to achieve 57–87% detection rates with zero false alarms on clean queries, while adding sub-millisecond latency per prompt. Critically, it works with any LLM provider without requiring model retraining, making it immediately actionable for teams running RAG systems, copilots, or tool-calling agents against untrusted inputs.

The article is written specifically for CACM's practitioner audience. Rather than focusing on novel algorithms, we synthesize industry patent strategies into a concrete deployment playbook, provide detailed implementation guidance (including pseudocode for the normalization pipeline), and quantify operational trade-offs between Production mode (zero false positives) and Monitoring mode (higher recall for threat intelligence). We answer the questions engineers actually ask before deploying security infrastructure: How vulnerable are current models? What detection approaches work? Do we need to tune thresholds? Can attackers evade with Unicode tricks? Is this fast enough? Every design choice is justified by empirical evidence from our systematic evaluation.

Key contributions suitable for CACM readers include:

- A model-agnostic firewall architecture deployable as middleware in front of existing LLM APIs, with quantified TPR/FAR metrics across multiple threat scenarios
- Synthesis of 18 industry patents (2023–2025) into five convergent defense patterns, showing our pipeline instantiates industry best practices
- Rigorous evaluation protocol spanning baseline vulnerability (400 attacks, 2 models), detector comparison, fusion strategies, threshold robustness, obfuscation testing (260 adversarial

benign queries), and generalization to 65 novel attacks

- Dual-mode deployment architecture (Production vs. Monitoring) with practical guidance on shadow logging, threshold validation, and continuous rule evolution
- Production-tested implementation with sub-millisecond latency and linear scaling characteristics

The work has been conducted independently at Videns, propelled by Cofomo, and has not been submitted elsewhere. All code, datasets, and experimental artifacts are available at `https://github.com/carlosdenner-videns/prompt-injection-cacm` to support reproducibility and practitioner adoption.

This manuscript aligns with CACM's mission to bridge research and practice in computing. By providing a deployable solution with transparent trade-offs, detailed implementation guidance, and systematic evaluation against realistic threats, we offer practitioners a concrete path to securing LLM applications today while the research community continues developing longer-term defenses.

Thank you for considering our submission. I look forward to your feedback and am happy to address any questions during the review process.

Sincerely,

Carlos Denner dos Santos
Research Scientist
Videns, propelled by Cofomo