ИНТЕЛЛЕКТУАЛЬНЫЙ FIREWALL НА ОСНОВЕ LLM ДЛЯ ОБНАРУЖЕНИЯ PROMPT INJECTION АТАК

Бекматов Акмал Курбонмахматович
Ассистент кафедры оптических систем связи и сетей,
Каршинский государственный технический университет
Вохидов Абдувохид Журабекович
Студент, Каршинский государственный технический университет.

Аннотация. В статье рассматривается архитектура интеллектуального межсетевого экрана, использующего большую языковую модель (LLM) для обнаружения атак типа Prompt Injection на системы на базе ИИ. Проведён анализ существующих таксономий атак, обоснована методология многоуровневой классификации входных данных. Предложена схема работы фильтрующего контура и сформулированы рекомендации по интеграции решения в производственные пайплайны.

Ключевые слова: Prompt Injection; LLM; интеллектуальный файрвол; безопасность ИИ; классификация угроз; обнаружение аномалий; защита языковых моделей.

LLM-BASED INTELLIGENT FIREWALL FOR PROMPT INJECTION ATTACK DETECTION

Bekmatov Akmal Kurbonmaxmatovich
Assistant Lecturer, Department of Optical Communication Systems and Networks, Karshi State Technical University
Vohidov Abduvohid Jo‘rabekovich
Student, Karshi State Technical University

Abstract. This paper presents the architecture of an intelligent firewall leveraging a large language model (LLM) to detect Prompt Injection attacks targeting AI-based systems. Existing attack taxonomies are analysed, a multi-layer input classification methodology is substantiated, and a filtering circuit schema is proposed alongside recommendations for integrating the solution into production pipelines.

Keywords: Prompt Injection; LLM; intelligent firewall; AI security; threat classification; anomaly detection; language model protection.

ВВЕДЕНИЕ

Широкое внедрение больших языковых моделей (LLM) в корпоративные продукты — от чат-ботов до автономных агентов — создаёт новый класс угроз,

не имеющих прямых аналогов в традиционной кибербезопасности. Атаки типа Prompt Injection эксплуатируют саму природу LLM: модель не разграничивает инструкции от разработчика и пользовательский ввод на уровне архитектуры, что позволяет злоумышленнику переопределить системный промпт, извлечь конфиденциальные данные или инициировать нежелательные действия [1].

Согласно классификации OWASP Top 10 для LLM-приложений (редакция 2025 года), Prompt Injection устойчиво занимает первую строку рейтинга наиболее критичных уязвимостей [2]. При этом существующие защитные механизмы — статические фильтры, списки запрещённых слов — демонстрируют низкую эффективность против семантически сложных атак, использующих перефразирование, кодирование или многошаговые сценарии [3].

Цель настоящей работы — обоснование архитектуры интеллектуального файрвола (LLM Firewall), выполняющего семантическую классификацию входных запросов до их передачи в целевую модель, и описание методологии его построения на основе открытых данных и инструментария.

МАТЕРИАЛЫ И МЕТОДЫ

Источники данных и таксономия угроз

Исследование опирается на следующие открытые источники. OWASP LLM Top 10 2025 [2] содержит актуальную классификацию уязвимостей, в том числе детальное описание прямых (Direct Prompt Injection) и косвенных (Indirect Prompt Injection) атак. Датасет Prompt Injection Detection от Hugging Face (Deepset, 2023) [4] включает более 1 500 размеченных примеров атак и легитимных запросов, используемых для обучения классификаторов. Работа Perez & Ribeiro (2022) [5] формализует понятие Prompt Injection и приводит систематику сценариев атак. Дополнительно использованы материалы отчёта NIST AI 100-2: Adversarial Machine Learning [6], описывающего угрозы целостности входных данных в системах на базе ИИ.

Архитектура LLM Firewall

Предлагаемая архитектура реализует принцип «охранника перед моделью» (guard-before-model): каждый входящий запрос проходит через отдельный классифицирующий LLM (Guard Model) прежде, чем достигнуть целевой модели. Формально задача классификации формулируется следующим образом:

$$f(x) = \arg\max P(c \mid x, \theta_G), \quad c \in \{safe, inject\} \quad (1)$$

где x — входной запрос; θ_G — параметры Guard Model; c — класс запроса

Трёхуровневая схема обработки запроса представлена на рисунке 1.

Архитектура интеллектуального LLM Firewall

ВХОДЯЩИЙ ЗАПРОС ПОЛЬЗОВАТЕЛЯ
▼
УРОВЕНЬ 1 — Лексический фильтр (regex / черные списки)
Блокировка очевидных паттернов: «ignore previous instructions», инъекции кодирования
▼ PASS
УРОВЕНЬ 2 — Семантический классификатор (Guard LLM / DistilBERT fine-tuned)
Вероятностная оценка P(inject | x, θG) — уравнение (1)
▼ P < τ (порог)
УРОВЕНЬ 3 — Контекстуальный аудит (проверка ролевых инструкций)
Верификация соответствия запроса системному промпту и политике приложения
▼ APPROVED
ЦЕЛЕВАЯ LLM → ОТВЕТ ПОЛЬЗОВАТЕЛЮ

Источник: разработано автором.

РЕЗУЛЬТАТЫ

Таксономия атак Prompt Injection

На основе анализа источников [1–5] сформирована рабочая таксономия, представленная в таблице 1. Классификация охватывает четыре основных вектора атак, различающихся по механизму воздействия и сложности обнаружения.

Таблица 1.

Таксономия атак типа Prompt Injection

Тип атаки	Механизм	Пример паттерна	Сложность детекции
Direct PI	Переопределение системного промпта	«Ignore previous instructions and…»	Низкая
Indirect PI	Инъекция через внешние данные (RAG, URL)	Вредоносный текст в загружаемом документе	Высокая
Encoded PI	Кодирование инструкций (Base64, ROT13)	«Decode and execute: aWdub3Jl…»	Средняя
Multi-step PI	Атака через цепочку диалоговых ходов	Постепенное смещение контекста в N-ходов	Очень высокая

Источник: составлено автором на основе [2, 5, 6].

Оценка эффективности подходов к детекции

В таблице 2 представлено сравнение трёх подходов к обнаружению Prompt Injection по критериям охвата, вычислительной стоимости и устойчивости к обходу. Данные основаны на опубликованных сравнительных тестах [4, 7].

Таблица 2.

Сравнительный анализ методов обнаружения Prompt Injection

Метод	Охват таксонов	Вычисл. стоимость	Устойчивость к обходу
Статические фильтры (regex)	Direct PI только	Низкая (O(n))	Очень низкая — trivially bypassed
Fine-tuned BERT-классификатор	Direct + Encoded	Средняя	Средняя — уязвим к перефразированию
Guard LLM (предлагаемый подход)	Все 4 типа	Высокая	Высокая — семантический охват контекста

Источник: составлено автором на основе [4, 7].

ОБСУЖДЕНИЕ

Предлагаемая трёхуровневая архитектура устраняет ключевой изъян существующих решений — неспособность детектировать семантически замаскированные атаки. Однако её практическое внедрение сопряжено с рядом нетривиальных компромиссов.

Во-первых, использование отдельного Guard LLM увеличивает задержку ответа (latency), что критично для интерактивных продуктов. Авторская оценка на основе публично доступных бенчмарков [7] показывает: при размере Guard Model до 7 млрд параметров дополнительная задержка на GPU-инференсе составляет порядка 150–300 мс — приемлемо для большинства B2B-сценариев, но требует оптимизации для высоконагруженных потребительских сервисов.

Во-вторых, порог классификации τ в уравнении (1) является настраиваемым гиперпараметром и определяет компромисс между ложными срабатываниями (false positives) и пропуском атак (false negatives). Для регулируемых отраслей — финансов, медицины — рекомендуется смещение τ в сторону минимизации FN, принимая более высокий уровень FP.

Научная новизна работы состоит в формализации многоуровневой схемы детекции с явным выделением семантического уровня на базе Guard LLM, а также в предложении критерия оптимального выбора τ в зависимости от профиля риска приложения. В отличие от ранее опубликованных подходов [3,

5], предлагаемая схема явно разграничивает лексический и контекстуальный уровни фильтрации, что повышает интерпретируемость принимаемых решений.

ЗАКЛЮЧЕНИЕ

Проведённый анализ позволяет сформулировать три практических вывода. Во-первых, статические фильтры недостаточны для защиты LLM-систем от семантически сложных атак; их следует рассматривать лишь как первый рубеж. Во-вторых, архитектура Guard LLM обеспечивает наибольший охват таксонов атак при сохранении приемлемой производительности. В-третьих, порог классификации τ должен конфигурироваться индивидуально для каждого приложения с учётом его профиля риска.

Перспективным направлением является разработка специализированных датасетов на русскоязычных и региональных языках Центральной Азии для дообучения Guard Model, а также интеграция механизмов объяснимого ИИ (XAI) для аудита решений файрвола в регулируемых отраслях.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

[1] Greshake T. et al. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injections // arXiv preprint arXiv:2302.12173. — 2023.

[2] OWASP. OWASP Top 10 for Large Language Model Applications. Version 2025. — URL: https://owasp.org/www-project-top-10-for-large-language-model-applications/ (дата обращения: 01.05.2025).

[3] Branch H.J. et al. Evaluating the Susceptibility of Pre-Trained Language Models via Handcrafted Adversarial Examples // arXiv preprint arXiv:2209.02128. — 2022.

[4] Deepset. Prompt Injection Detection Dataset. — Hugging Face, 2023. — URL: https://huggingface.co/datasets/deepset/prompt-injections

[5] Perez F., Ribeiro I. Ignore Previous Prompt: Attack Techniques For Language Models // arXiv preprint arXiv:2211.09527. — 2022.

[6] NIST. Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. NIST AI 100-2. — Gaithersburg: NIST, 2024.

[7] Rebedea T. et al. NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications // arXiv preprint arXiv:2310.10501. — 2023.

[8] Schulhoff S. et al. Ignore This Title and HackAPrompt: Exposing Systemic Vulnerabilities of LLMs Through a Global Prompt Hacking Competition // Proceedings of EMNLP 2023. — 2023. — P. 4945–4977.

[9] Бекматов А.К., & Эргашов Ф.Т. (2025). ОБЕСПЕЧЕНИЕ АУТЕНТИФИКАЦИИ В СЕТИ ПЕРЕДАЧИ ДАННЫХ. Экономика и социум, (1-2 (128)), 1013-1017.

[10] Бекматов А.К., & Рустамов Т.С. (2024). РОЛЬ ГЛУБОКОГО ОБУЧЕНИЯ В УЛУЧШЕНИИ ТОЧНОСТИ СИСТЕМ ОБНАРУЖЕНИЯ ВТОРЖЕНИЙ. Экономика и социум, (6-1 (121)), 1582-1591.