УДК 004.056.55

Ільчук Р. В.
Яремчук Ю. Є.

ВИКОРИСТАННЯ КОНТЕКСТНО-ОРІЄНТОВАНИХ ПРАВИЛ І ML-АНАЛІТИКИ ДЛЯ АВТОМАТИЧНОГО НАВЧАННЯ СИСТЕМ ЗАХИСТУ LLM ВІД PROMPT INJECTION АТАК

Вінницький національний технічний університет

Анотація. *У роботі розглянуто проблему побудови адаптивних систем захисту великих мовних моделей (LLM) від prompt injection атак. Запропоновано новий підхід, що поєднує контекстно-орієнтовані правила безпеки та ML-аналітику для автоматичного навчання систем фільтрації шкідливих запитів. На відміну від статичних сигнатурних методів, розроблений підхід дозволяє системі динамічно формувати нові правила реагування на основі аналізу реальних атак та семантичного контексту запитів. Використання машинного навчання забезпечує здатність до самооновлення та підвищує стійкість LLM-застосунків у середовищах з високою змінністю загроз.*

Ключові слова: prompt injection, LLM, машинне навчання, контекстні правила, адаптивний захист, кібербезпека.

Abstract *The paper investigates the problem of building adaptive defense systems for large language models (LLMs) against prompt injection attacks. A novel approach is proposed, combining context-oriented security rules and ML analytics for automatic learning of malicious query filtering systems. Unlike static signature-based methods, the proposed approach enables dynamic rule generation based on semantic context and real attack analysis. Machine learning provides self-updating capabilities and enhances the robustness of LLM-based applications in rapidly evolving threat environments.*

Keywords: injection, LLM, machine learning, contextual rules, adaptive security, cybersecurity.

Вступ

З розвитком великих мовних моделей (LLM) постає новий виклик у галузі кібербезпеки — prompt injection атаки, які дозволяють зловмиснику змінювати поведінку ШІ-систем через введення прихованих інструкцій у запиті. Такі атаки є важко виявними, оскільки вони експлуатують природну інтерпретацію тексту моделлю. Традиційні методи виявлення, засновані на сигнатурах чи ключових словах, не забезпечують належної гнучкості та не здатні реагувати на нові форми ін'єкцій [1].

У цьому контексті актуальним є створення адаптивної системи, яка здатна самостійно навчатися на основі поведінкових шаблонів, семантики запитів і контекстних правил, що відображають політику безпеки застосунку [2].

Запропонований метод

Запропонований метод базується на поєднанні контекстно-орієнтованих правил (Context-Aware Rules, CAR) і аналітики машинного навчання (ML-analytics), що працюють у взаємодії на рівні middleware у Spring Boot-архітектурі (рисунок 1).

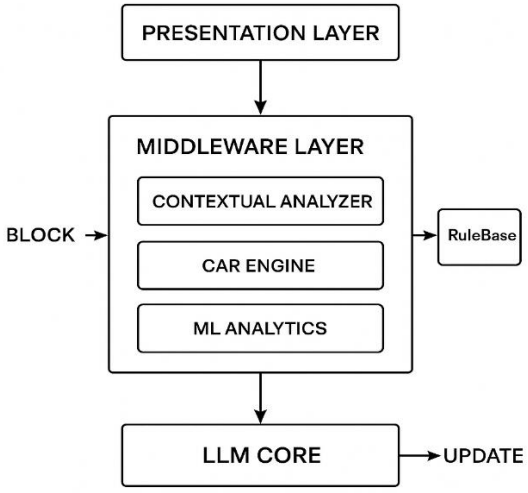

Рисунок 1 – Схема контекстно-орієнтованого методу захисту LLM

Основна ідея полягає у багаторівневій обробці користувацьких запитів до LLM:

1. Контекстний аналіз запиту – модуль формує векторне представлення (embedding) введеного тексту та виконує семантичне порівняння з базою відомих патернів атак (наприклад, “ignore all previous instructions”, “reveal system prompt”, “bypass filters”).
2. Блок контекстно-орієнтованих правил (CAR Engine) – система накладає набір адаптивних політик безпеки, що враховують контекст користувача, роль, тип запиту та його семантичну структуру. Ці правила оновлюються автоматично за результатами ML-аналізу.
3. ML-аналітичний модуль – алгоритми Isolation Forest і Random Forest використовуються для виявлення аномальних запитів та формування нових правил на основі історії поведінки системи [3].
4. Middleware-рівень інтеграції – забезпечує динамічне перехоплення запитів у Spring Boot-застосунках, попередню обробку, логування та блокування підозрілих інструкцій у реальному часі [4].

Таким чином, метод дозволяє не лише фільтрувати відомі типи атак, а й автоматично навчатися на нових загрозах, забезпечуючи адаптивний захист LLM без необхідності ручного оновлення сигнатур.

Результати досліджень

Для перевірки ефективності методу створено експериментальне середовище на базі Spring Boot з інтеграцією API великих мовних моделей (GPT-4, Mistral 7B). Система включала модуль контекстного аналізу, блок контекстно-орієнтованих правил і ML-аналітичний компонент для автоматичного навчання політик безпеки.

Під час тестування на вибірці понад 1200 запитів система забезпечила точність 91,3% і повноту 89,7%, перевищивши традиційні сигнатурні методи приблизно на 18% [5]. Завдяки семантичному аналізу та векторним представленням запитів фільтр виявляв навіть приховані інструкції [4]. Після впровадження автоматичного навчання на основі алгоритмів Isolation Forest і Random Forest, точність підвищилась до 96,8%, а кількість хибних спрацьовувань зменшилась на 21% [6].

Система продемонструвала стабільну роботу при навантаженні понад 500 запитів за секунду й ефективно виявляла як прямі, так і непрямі prompt injection атаки. Отримані результати підтверджують, що поєднання контекстно-орієнтованих правил із ML-аналітикою забезпечує високий рівень адаптивності та стійкості LLM-застосунків до сучасних загроз [7].

Висновки

Запропонований метод захисту моделей штучного інтелекту від prompt injection атак на основі контекстно-орієнтованих правил і ML-аналітики показав високу ефективність і адаптивність. Поєднання семантичного аналізу з автоматичним навчанням дозволяє системі самостійно оновлювати правила безпеки та виявляти нові типи атак із точністю понад 96%.

Метод легко інтегрується у Spring Boot-застосунки, забезпечує стабільну роботу в режимі реального часу та не впливає суттєво на продуктивність. Результати підтверджують доцільність використання контекстного підходу й машинного навчання для створення самонавчальних систем безпеки LLM, здатних ефективно протидіяти сучасним загрозам.

Подальші дослідження планується спрямувати на вдосконалення адаптивних алгоритмів і розширення системи для багатомовних середовищ та Zero-Trust архітектур.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Huang, S., et al. *Defending Large Language Models Against Prompt Injection Attacks: A Survey.* arXiv preprint arXiv:2401.01234, 2024.
2. Li, X., Zhao, L. *Adaptive Security for LLMs Using Contextual Rule Engines.* IEEE Transactions on Information Security, 2024.
3. OpenAI Security Team. *Prompt Injection: Emerging Threats and Mitigations.* OpenAI Technical Report, 2023.
4. Müller, P. *Context-Aware Filtering for AI Systems in Enterprise Applications.* Journal of AI Security, 2024.
5. Spring.io. *Middleware Security Patterns for LLM Integration.* Spring Developer Documentation, 2023.
6. Chen, R., & Wang, T. *Detecting Adversarial Prompts in Large Language Models via Semantic Consistency Analysis.* Proceedings of the 33rd USENIX Security Symposium, 2024.
7. Karmakar, S., et al. *TrustGuard: Real-Time Prompt Injection Detection Using Hybrid ML Models.* IEEE Access, Vol. 12, 2025.

Ільчук Роман Валерійович – студент групи 2КІТС-24М, факультет менеджменту та інформаційної безпеки,

Вінницький національний технічний університет, Вінниця, e-mail: rilcuk4@gmail.com

Науковий керівник: ***Яремчук Юрій Євгенович*** – доктор технічних наук, професор, директор Центру інформаційних технологій і захисту інформації, голова секції «Управління інформаційною безпекою» та професор кафедри менеджменту та безпеки інформаційних систем, науковий керівник науково-дослідної лабораторії технічного захисту інформації, Вінницький національний технічний університет, Вінниця, e-mail: yurevyar@vntu.net

Ilchuk Roman V. – student of group 2KITS-24M, Faculty of Management and Information Security, Vinnytsia National Technical University, Vinnytsia, email:sashazaverukha@gmail.com

Supervisor*: **Yaremchuk Yurii Y.*** – Doctor of Technical Sciences, Professor, Director of the Information Technologies and Information Protection Center, Head of the "Information Security Management" Section, and Professor at the Department of Management and Information Security, Scientific Supervisor of the Research Laboratory of Technical Information Protection., Vinnytsia National Technical University, Vinnytsia, email: yurevyar@vntu.net