УДК 004.89

МЕТОДИ ВИЯВЛЕННЯ ТА ПРОТИДІЇ АТАКАМ PROMPT INJECTION У СИСТЕМАХ НА ОСНОВІ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ

Лавська В.В., Погурська М.М.
e-mail: valeriia.lavska@nure.ua, mariia.yerovenko@nure.ua
Харківський національний університет радіоелектроніки, каф. ШІ
м. Харків, Україна

This work analyzes the problem of prompt injection attacks in LLM-based systems and reviews existing approaches to their detection and mitigation. The research highlights the importance of developing reliable security mechanisms for modern AI-driven systems. Large language models (LLMs) have become one of the most influential technologies in modern artificial intelligence and data mining. They are widely used in intelligent assistants, chatbots, document analysis systems and decision support applications. However, the rapid adoption of LLM-based systems introduces new security challenges. One of the most critical threats is prompt injection, where maliciously crafted prompts manipulate the behavior of a language model and bypass system restrictions.

Останніми роками великі мовні моделі (Large Language Models, LLM) стали однією з ключових технологій сучасного штучного інтелекту та інтелектуального аналізу даних. Завдяки використанню архітектури трансформерів такі моделі демонструють високі результати у задачах обробки природної мови, автоматичного перекладу, аналізу текстових документів, генерації природної мови та автоматизації взаємодії з користувачами [1].

Одним із важливих напрямів розвитку LLM є інтеграція з підсистемами пошуку та аналізу даних. Зокрема, значного поширення набули архітектури Retrieval-Augmented Generation (RAG), які поєднують можливості мовних моделей із механізмами пошуку релевантної інформації у зовнішніх джерелах (корпоративні бази знань, документні сховища, веб-ресурси тощо). Такий підхід дозволяє підвищити точність і обґрунтованість відповідей моделі за рахунок використання актуального контексту та залучення зовнішніх знань [2]. Водночас саме інтеграція із зовнішніми джерелами та зростання складності сценаріїв взаємодії з користувачем підвищують вимоги до безпеки LLM-систем.

Стрімке впровадження великих мовних моделей у реальні інформаційні системи супроводжується появою нових кібербезпекових ризиків. На відміну від традиційних програмних систем, LLM сприймають текстові інструкції як частину контексту обробки запиту. Це означає, що спеціально сформульовані команди або текстові маніпуляції можуть змінювати поведінку моделі та впливати на її рішення. У результаті

виникає можливість здійснення атак, спрямованих на обхід обмежень, порушення політик безпеки, витік конфіденційної інформації або виконання небажаних дій [3].

Однією з найбільш поширених і критичних загроз для систем на основі LLM є атаки Prompt Injection. Такі атаки полягають у навмисному формуванні запитів або інструкцій, що змінюють поведінку мовної моделі, змушують її ігнорувати встановлені правила або підштовхують до небезпечних дій. Потенційні наслідки включають розкриття конфіденційних даних, порушення політик безпеки, отримання доступу до внутрішніх інструкцій системи, а також генерацію небезпечного чи забороненого контенту [3]. Важливою особливістю Prompt Injection є те, що атаки можуть реалізовуватися на рівні текстових інструкцій без необхідності прямого доступу до програмного коду системи, що ускладнює їх виявлення та запобігання у практичних сценаріях.

Зазвичай виділяють два основні типи атак Prompt Injection: direct та indirect. У випадку direct prompt injection шкідлива інструкція безпосередньо включається до запиту користувача, наприклад, через спроби примусити модель ігнорувати попередні правила, розкрити системні повідомлення або видати конфіденційні дані. Натомість indirect prompt injection характерний для RAG-систем і пов’язаний з тим, що шкідливі інструкції можуть бути приховані у зовнішніх джерелах інформації, які система підтягує як контекст для відповіді. У такому випадку модель може інтерпретувати частину retrieved-тексту не як інформаційний зміст, а як команду, що призводить до небажаної поведінки. Це створює додаткову поверхню атаки та підвищує ризики для систем, що працюють з великими корпусами документів.

Вразливість LLM до Prompt Injection пов’язана з фундаментальною особливістю їх роботи: модель обробляє усі вхідні дані як частину єдиного контексту та не завжди здатна надійно відокремлювати системні інструкції від користувацьких. Зловмисники можуть використовувати різні техніки для обходу обмежень, зокрема: примус моделі ігнорувати попередні інструкції; спроби отримати системний prompt; запити на розкриття конфіденційної інформації; нав’язування небезпечних цілей; комбінування багатокрокових інструкцій у межах одного або кількох запитів. З огляду на практичну значущість проблеми атаки Prompt Injection включено до переліку найбільш критичних ризиків у рамках ініціативи OWASP Top 10 for Large Language Model Applications [4].

Для зменшення ризиків Prompt Injection у сучасних дослідженнях і практичних впровадженнях пропонуються різні підходи до захисту LLM-систем. Одним із базових підходів є rule-based фільтрація, яка передбачає пошук небезпечних ключових фраз та патернів у запитах користувача (або у retrieved-контексті) з метою блокування потенційно шкідливих інструкцій. Проте ефективність таких методів обмежена, оскільки

зловмисники можуть модифікувати формулювання атак, використовувати перефразування, приховані інструкції або контекстні прийоми, які складно описати простими правилами [5]. Іншим підходом є ізоляція контексту, що зменшує ймовірність того, що шкідлива інструкція буде сприйнята як частина системного керування моделлю.

Перспективним напрямом є використання методів машинного навчання та інтелектуального аналізу текстових даних для автоматичного виявлення атак Prompt Injection. У цьому випадку будуються класифікаційні моделі, які аналізують текст запиту (або контекст) та визначають, чи містить він ознаки шкідливого впливу. Такі рішення можуть використовувати різні підходи, включно з векторними представленнями тексту (embeddings), трансформерними моделями типу BERT, а також класичними алгоритмами класифікації текстових даних [5]. З позиції Data Mining це дозволяє виявляти складніші шаблони атак, аналізувати закономірності шкідливих інструкцій, підвищувати узагальнювальну здатність детекторів та покращувати практичну стійкість систем безпеки.

У межах подальших досліджень планується розроблення методу автоматичного виявлення атак Prompt Injection на основі аналізу текстових запитів із використанням моделей машинного навчання. Запропонований підхід передбачає застосування методів інтелектуального аналізу текстових даних для класифікації запитів користувача та виявлення потенційно шкідливих інструкцій. Очікується, що використання трансформерних моделей та векторних представлень тексту дозволить підвищити точність виявлення атак і покращити рівень безпеки систем на основі великих мовних моделей.

Список використаних джерел:

1. Vaswani A. Attention Is All You Need. arXiv.org. DOI: https://doi.org/10.48550/arXiv.1706.03762.

2. Lewis P. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv.org. DOI: https://doi.org/10.48550/arXiv.2005.11401.

3. Greshake K. Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. arXiv.org. DOI: https://doi.org/10.48550/arXiv.2302.12173.

4. OWASP Top 10 for Large Language Model Applications | OWASP Foundation. OWASP. URL: https://owasp.org/www-project-top-10-for-large-language-model-applications/ (date of access: 05.03.2026).

5. Zhu K. PromptRobust: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts. arXiv.org. DOI: https://doi.org/10.48550/arXiv.2306.04528.