

ПОЗАЕКРАННІ ЕКСПЛОЙТИ: ЯК PROMPT-ІН'ЄКЦІЇ ПЕРЕТВОРЮЮТЬ LLM-АГЕНТІВ НА НЕВИДИМИХ ХАКЕРІВ

Вінницький національний технічний університет

Анотація

Робота присвячена дослідженню нового класу кібератак на основі prompt-ін'єкцій у автономні агенти штучного інтелекту. Показано, що делегування моделям великих мов (LLM) функцій взаємодії з реальними системами створює принципово нові вектори зламу, які виходять за рамки класичних технічних експлоїтів. Проаналізовано механізми атак через prompt-ін'єкції, ланцюги постачання моделей, AI-підсилене шкідливе програмне забезпечення та шляхи створення захисту на рівні смислових інструкцій.

Ключові слова: prompt-ін'єкції, LLM-агенти, кібербезпека, AI supply chain, jailbreak, автономні системи штучного інтелекту.

Abstract

This paper is devoted to the study of a new class of cyberattacks based on prompt injections targeting autonomous artificial intelligence agents. It is shown that delegating real-world interaction tasks to large language models (LLMs) creates fundamentally new attack vectors that extend beyond traditional technical exploits. The paper analyzes attack mechanisms involving prompt injections, model supply chains, AI-enhanced malware, and approaches to developing defenses at the level of semantic instructions.

Keywords: prompt injection, LLM agents, cybersecurity, AI supply chain, jailbreak, autonomous artificial intelligence systems.

Вступ

Сучасна кібербезпека переживає перехід від захисту портів, процесів і пакетів до захисту шару смислів, де рішення приймають не люди, а штучний інтелект. З появою автономних агентів на базі великих мовних моделей (LLM), здатних виконувати дії в браузері, електронній пошті та корпоративних системах, атаки через prompt-ін'єкції стають новим класом позаекранних експлоїтів. У 2025–2026 роках фокус зміщується від абстрактних «AI-ризиків» до конкретних інцидентів: витоків даних через генеративні сервіси, зловживань автономними агентами, компрометації ланцюгів постачання моделей. Тому актуальною стає систематизація механізмів атак через prompt-ін'єкції на LLM-агентів, аналіз ризиків безпеки ланцюгів постачання моделей штучного інтелекту та формулювання напрямів створення захисту на рівні семантичних інструкцій. Дослідження базується на аналізі звітів World Economic Forum Global Cybersecurity Outlook 2026 [1], OWASP Top 10 for LLM Applications 2025, матеріалів конференцій з AI-безпеки, а також документованих випадків атак через prompt-ін'єкції (CVE-2025-54794, випадки експлуатації AI-браузерів). Застосовано методи порівняльного аналізу традиційних і AI-орієнтованих векторів атак, класифікацію prompt-ін'єкційних технік за типами та дослідження ланцюгів постачання моделей.

Результати дослідження

Prompt-ін'єкції як новий клас експлоїтів. Характерна особливість сучасних атак полягає в тому, що мішенню стає не сама мовна модель, а її контекст і зв'язки з іншими сервісами. Агент, який читає пошту, переходить за посиланнями, генерує відповіді у браузері та викликає внутрішні API, мимоволі перетворюється на «скрипт-кідді з дипломом», якщо його системні підказки можна переписати через зовнішній контент [2]. Описаний у 2025 році випадок демонструє перехоплення облікового запису одним кліком: шкідливий коментар у Reddit примушував AI-браузер прочитати одноразовий пароль із пошти жертви й відправити його назад у тред.

Ланцюги постачання моделей. Якщо раніше supply chain-ризик означали скомпрометовані бібліотеки та CI/CD, то тепер у цю картину додаються отруєні датасети, моделі з прихованими тригерами та небезпечні плагіни до LLM-платформ [3]. Вставка бекдору на етапі тренування створює модель, яка виглядає нормально під час звичайного тестування, але радикально змінює поведінку при спеціальному текстовому маркері в запиті. У поєднанні з автономними діями це відкриває шлях до атак, де один рядок у PDF запускає ланцюжок операцій із реальними грошима, контрактами чи доступами.

AI-підсилені традиційні атаки. Взаємодія AI-загроз із класичними сценаріями – фішингом, вимагальницьким ПЗ, компрометацією бізнес-пошти – створює вибух персоналізованих атак [4]. Моделі генерують повідомлення, максимально схожі на стиль конкретної людини, автоматично аналізують відповіді жертви й адаптують сценарій у реальному часі. AI-підсилений ransomware в режимі реального часу шукає вразливості, відтинає резервні копії, обирає об'єкти для шантажу та формує меседжі для конкретних менеджерів. Межа між людськими та машинними атаками практично стирається.

Напрями захисту. Реакція галузі зміщається від панічних заборон до формування повноцінного «AI-циклжиттєвого» захисту – від вимог до даних і моделей до експлуатації та виведення з експлуатації. Регулятори пропонують перші рамки для безпечного використання генеративного AI: ізоляція агентів, валідація дій, моніторинг ланцюгів постачання моделей, управління ризиками prompt-ін'єкцій. Досвід показує, що технічних патчів і фільтрів недостатньо – потрібен новий спосіб мислення, де безпека на рівні протоколів доповнюється безпекою на рівні смислів, інструкцій та взаємодій людина–AI–система[5].

Висновки

Prompt-ін'єкції та атаки на AI-агентів є першою хвилею семантичних експлоїтів, де «вразливість» живе не лише в байтах, а й у способі делегування моделі прийняття рішень. Для кібербезпеки це шанс сформувати нову дисципліну – безпеку автономних рішень, що поєднує традиційний AppSec, захист ланцюга постачання та розуміння когнітивних меж самих моделей. Подальші дослідження мають зосередитися на розробці методів верифікації дій агентів, детекції непрямих prompt-ін'єкцій і створенні стандартів безпеки для ланцюгів постачання LLM.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. World Economic Forum. Global Cybersecurity Outlook 2026. URL: <https://www.weforum.org/publications/global-cybersecurity-outlook-2026/>
2. Hackers-Arise. What Will Be Key Cybersecurity Issues in 2026? URL: <https://hackers-arise.com/what-will-be-key-cybersecurity-issues-in-2026/>
3. Cybersecurity Dive. 5 cybersecurity trends to watch in 2026. URL: <https://www.cybersecuritydive.com/news/5-cybersecurity-trends-2026/810354/>
4. CYE. 2026: The Year Cybersecurity Crosses Every Line. 2025. URL: <https://cyesec.com/blog/2026-the-year-cybersecurity-crosses-every-line>
5. Capgemini. Trends in Cybersecurity 2025/2026. 2025. URL: <https://www.capgemini.com/>

Каратнюк Анна Русланівна – студентка групи ЗКІТС-246, факультет менеджменту та інформаційної безпеки, Вінницький національний технічний університет, м. Вінниця, e-mail: annakaradniuk505@gmail.com

Салієва Ольга Володимирівна – доктор філософії з кібербезпеки (PhD), доцент кафедри менеджменту та безпеки інформаційних систем, e-mail: salieva8257@vntu.edu.ua;

Saliieva Olha V. – Doctor of Philosophy in Cybersecurity (PhD), Associate Professor of the Department of Management and Information Systems Security, e-mail: salieva8257@vntu.edu.ua;

Anna Karatnyuk – student in Faculty of Management and Information Security, Vinnytsia National Technical University, Vinnytsia, e-mail: annakaradniuk505@gmail.com