УДК 004.89.056.53

ЕКСПЛУАТАЦІЯ ВРАЗЛИВОСТЕЙ ШТУЧНОГО ІНТЕЛЕКТУ ТА СПОСОБИ ЇХ ЗАПОБІГАННЯ

Руденко Б.В.
e-mail: bohdan.rudenko@nure.ua
Науковий керівник – канд. техн. наук, ст. викл. Кобилін І.О.
Харківський національний університет радіоелектроніки, каф. ІНФ
м. Харків, Україна

The growing integration of AI into critical infrastructure introduces new cybersecurity threats targeting large language models (LLMs) and agentic AI systems. Threat actors now embed LLMs directly into malware for real-time code generation and self-obfuscation, while autonomous AI agents can perform up to 90% of intrusion operations with minimal human involvement. This paper classifies AI vulnerabilities into model-level, data-level, prompt-based and orchestration-layer categories, examines social engineering techniques for bypassing model safeguards, and proposes mitigation measures including secure AI development lifecycles, risk management frameworks, monitoring with kill-switch mechanisms, and defensive AI-driven threat detection.

Штучний інтелект швидко стає невід'ємною частиною державних інформаційних систем, бізнес-процесів і об'єктів критичної інфраструктури, тому його компрометація безпосередньо впливає на національну безпеку. У 2025 році Google Threat Intelligence Group зафіксувала перехід від використання ШІ як «розумного помічника» до вбудовування великих мовних моделей безпосередньо в шкідливе ПЗ для генерації коду, самообфускації та динамічної зміни поведінки під час атаки. Паралельно звіт Anthropic описує першу задокументовану кампанію кібершпигунства, де агентна модель забезпечувала до 80–90% тактичної роботи – від розвідки до ексфільтрації даних – тоді як люди брали участь переважно на рівні стратегічних рішень.

Ці приклади свідчать, що вразливості ШІ-систем мають розглядатися не як окремі помилки реалізації, а як комплексна модель загроз, що охоплює модель, дані, промпти, оркестрацію інструментів і взаємодію з користувачем. Атаці підлягає весь ланцюжок «користувач–промпт–модель–інструменти–інфраструктура», а наслідки можуть включати як цілеспрямований витік конфіденційної інформації, так і масштабні порушення роботи сервісів.

Можна виділити три основні класи вразливостей. Перший – модельні та архітектурні вразливості: здатність моделі виконувати інструкції, що обходять формальні політики безпеки, схильність до «галюцинацій», а також можливість використання моделей як «рушіїв мутації», коли шкідливий код регулярно переписується LLM для ускладнення виявлення. Другий клас – вразливості даних і контексту: отруєння тренувальних датасетів, ін'єкція промптів через зовнішні джерела, маніпуляції з правами доступу та компрометація API-ключів. Третій – інтеграційні та оркестраційні вразливості, що

виникають, коли агентні моделі отримують доступ до розгалуженого набору MCP-інструментів: сканерів, браузерної автоматизації, баз даних тощо.

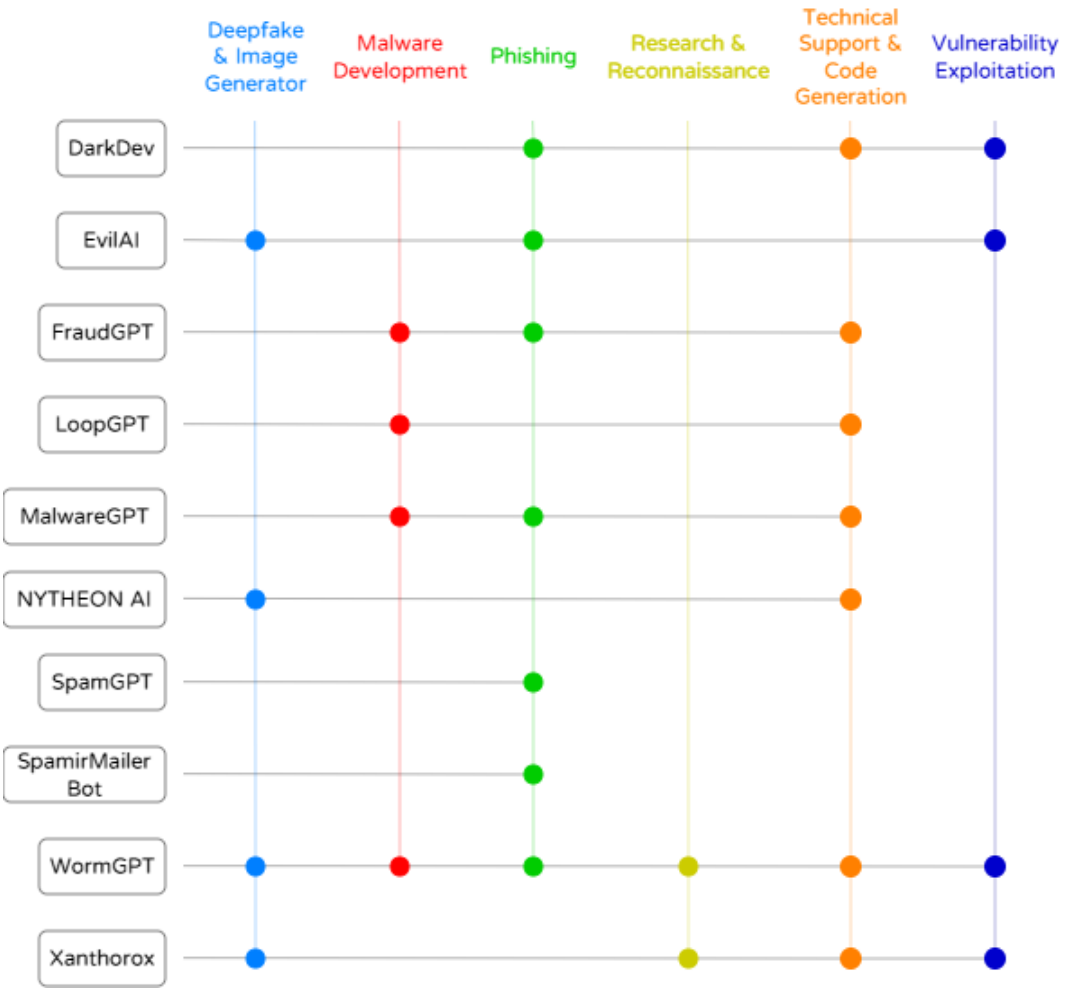

Рисунок 1 – Можливості відомих інструментів та сервісів штучного інтелекту, рекламованих на англомовних та російськомовних підпільних форумах

У відкритих джерелах описано низку AI-орієнтованих кампаній. Звіт Google наводить сімейства шкідливого ПЗ FRUITSHELL, PROMPTFLUX, PROMPTSTEAL, PROMPTLOCK і QUIETVAULT, що демонструють вбудовування LLM у цикл атаки. Зокрема, PROMPTFLUX використовує API моделі для регулярного переписування власного VBScript-коду з метою обходу антивірусів. PROMPTSTEAL, застосований російським APT28 проти України, використовував LLM як «генератор системних команд»: шкідливий модуль відправляв до моделі запит на збирання даних, отримував однорядкові команди, виконував їх локально та передавав результати на C2-сервер. Це фактично виносить ключову логіку атаки за межі аналізованого бінарного файлу.

Кампанія GTG-1002, описана Anthropic, демонструє ще вищий рівень автономності: агентна система самостійно виконувала масову розвідку

близько тридцяти цілей, виявляла вразливості, генерувала експлойти, будувала карти внутрішніх мереж і вилучала конфіденційні дані. Окремим вектором є маніпуляція промптами: загрозливі актори прикидаються учасниками CTF-змагань, студентами або фахівцями з кібербезпеки, що дозволяє оминути базові фільтри моделей. Така «соціальна інженерія ШІ» показує, що моделі потребують як технічних, так і процедурних запобіжників.

Для зменшення ризиків необхідний поєднаний підхід. По-перше, впровадження безпечного життєвого циклу розробки ШІ (secure AI SDLC), що включає загрозомоделювання, тестування на стійкість до промпт-ін'єкцій і «red teaming» відповідно до NIST AI RMF. По-друге, моніторинг використання моделей і механізми швидкого втручання: блокування облікових записів, відкликання ключів, оновлення класифікаторів. Практика Google та Anthropic показує, що після виявлення зловживань компанії оновлюють захисні системи так, щоб подібні сценарії відхилялися автоматично. По-третє, відповідно до рекомендацій ENISA, в агентних системах необхідно обмежувати доступ за принципом найменших привілеїв і впроваджувати незалежні контрольні механізми. Нарешті, перспективним напрямом є використання самого ШІ для оборони: автоматизований аналіз журналів безпеки, виявлення аномалій та побудова систем раннього попередження про промпт-ін'єкції.

Експлуатація вразливостей штучного інтелекту стала реальним інструментом державних та кримінальних зловмисників. Найбільшу небезпеку становить поєднання агентних можливостей моделей, широкого доступу до зовнішніх інструментів і здатності зловмисників маніпулювати контекстом промптів. Необхідно розвивати фреймворки безпечної розробки ШІ, посилювати моніторинг і механізми реагування, а також системно використовувати AI-технології для захисту.

Список використаних джерел:

1. Google Threat Intelligence Group. Advances in Threat Actor Usage of AI Tools. November 2025.

2. Anthropic. Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign. Full Report. November 2025.

3. National Institute of Standards and Technology (NIST). Artificial Intelligence Risk Management Framework (AI RMF 1.0) January 2023.