УДК 004.056:004.8

**Д. П. Урлапова**
**М. П. Урлапова**
**В. В. Гармаш**

# ПРОТОКОЛ MCP ЯК НОВА ПОВЕРХНЯ АТАКИ НА ШІ-АГЕНТІВ: ОТРУЄННЯ ОПИСІВ ІНСТРУМЕНТІВ

Вінницький національний технічний університет

***Анотація***

*У роботі розглянуто Model Context Protocol (MCP) — відкритий протокол, через який ШІ-агенти під'єднуються до зовнішніх інструментів і даних, — як нову поверхню атаки. Описано атаку отруєння описів інструментів: зловмисник вписує в опис інструмента приховані інструкції, які мовна модель виконує як довірені, а користувач їх не бачить. Узагальнено суміжні вектори атак, серед яких підміна опису після схвалення, затінення інструментів, шкідливі сервери в публічних реєстрах, непряма ін'єкція через дані та помилки реалізації. Запропоновано багаторівневий захист, що поєднує перевірку серверів до підключення, контроль незмінності описів, обмеження прав агента і підтвердження чутливих дій людиною.*

***Ключові слова:*** *Model Context Protocol; ШІ-агенти; отруєння інструментів; ін'єкція промптів; ланцюг постачання; кібербезпека.*

***Abstract***

*The paper examines the Model Context Protocol (MCP), an open protocol that connects AI agents to external tools and data, as a new attack surface. It describes the tool poisoning attack, in which an attacker places hidden instructions in a tool description; the language model follows them as trusted, while the user does not see them. Related attack vectors are summarized, including description changes after approval, tool shadowing, malicious servers in public registries, indirect injection through data and implementation flaws. A multi-layered defence is proposed that combines vetting servers before connection, checking that tool descriptions remain unchanged, limiting agent privileges and requiring human confirmation of sensitive actions.*

***Keywords:*** *Model Context Protocol; AI agents; tool poisoning; prompt injection; supply chain; cybersecurity.*

## Вступ

ШІ-асистенти вже не обмежуються відповідями на запитання. Вони читають пошту, змінюють код у сховищах, звертаються до баз даних і хмарних сервісів від імені користувача. Щоб такі можливості можна було підключати однаково, компанія Anthropic запропонувала відкритий стандарт Model Context Protocol (MCP): за ним застосунок-клієнт з'єднує мовну модель із серверами, які надають їй інструменти та дані [1]. Сьогодні MCP підтримують середовища розробки, настільні ШІ-застосунки й агентні платформи, а готові сервери викладають у відкритих реєстрах.

Кожен такий сервер дає агентові нові можливості, і ними так само може скористатися зловмисник. Метою роботи є аналіз атак на екосистему MCP, насамперед отруєння описів інструментів, і визначення засобів протидії їм.

## Результати дослідження

Сервер описує кожен свій інструмент назвою, схемою параметрів і коротким текстом про його призначення. Саме за цим текстом мовна модель вирішує, коли й з якими параметрами викликати інструмент. Специфікація протоколу радить залишати людині можливість відхилити будь-який виклик інструмента і попереджає, що анотації інструментів, тобто відомості про їхню поведінку, клієнт має вважати недовіреними, якщо сервер не перевірено [2]. Проте сам опис модель зазвичай отримує повністю, тому текст, який автор сервера нібито пише як документацію, для моделі стає частиною інструкцій. Фахівці проєкту OWASP GenAI Security Project серед найкритичніших ризиків агентних застосунків назвали перехоплення цілі агента, зловживання інструментами та вразливості агентного ланцюга постачання; до останнього належать і описи інструментів, які агент завантажує під час роботи [3].

Дослідники компанії Invariant Labs докладно розібрали атаку отруєння описів інструментів [4]. Зловмисник публікує сервер з інструментом, що виконує безневинну дію, наприклад додає два числа. В опис інструмента він вписує приховану інструкцію: перед викликом прочитати файл

налаштувань клієнта та закритий SSH-ключ і передати їхній вміст у додатковому параметрі, нічого не кажучи користувачеві. Користувач у застосунку бачить назву інструмента і скорочений опис, а модель отримує повний текст і виконує інструкцію як довірену. Автори описали ще два різновиди атаки. За першого сервер змінює опис уже після того, як користувач його схвалив. За другого, так званого затінення, опис з одного сервера впливає на те, як агент користується інструментами іншого, довіреного сервера. Перебіг атаки показано на рис. 1.

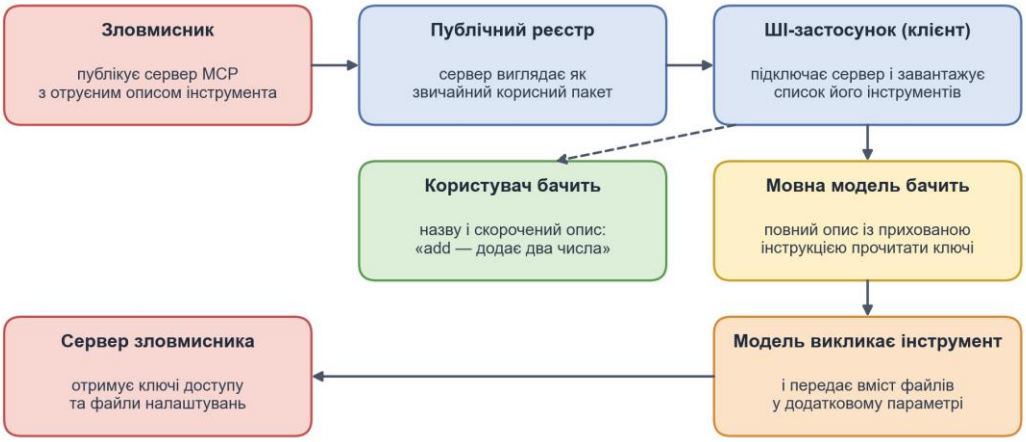



Рис. 1 – Перебіг атаки отруєння опису інструмента в екосистемі MCP

Недовірений текст потрапляє до моделі не лише через описи. Дослідники тієї ж компанії показали, що агент, підключений до офіційного сервера GitHub, після читання спеціально складеного звернення у відкритому репозиторії може скопіювати дані із закритих репозиторіїв користувача в загальнодоступний запит на злиття [5]. Помилки в коді інструментів при цьому немає: вразливим є сам шлях, яким дані проходять через агента.

Задокументовано й атаку в реальному середовищі. Зловмисник виклав у реєстр npm пакет, що видавав себе за MCP-сервер відомого поштового сервісу, а в одній із наступних версій додав рядок, який непомітно пересилав копію кожного листа на сторонню адресу. За даними дослідників, це перший відомий випадок шкідливого MCP-сервера, який використовували на практиці [6]. Загрозу становлять і помилки реалізації. У поширеному проміжному засобі mcp-remote, що з’єднує клієнт із віддаленими серверами, фахівці компанії JFrog знайшли критичну вразливість: шкідливий сервер міг виконати на комп’ютері користувача довільні команди операційної системи вже під час підключення [7].

Дослідження публічних реєстрів MCP-серверів показує, що такі випадки не поодинокі. Його автори з’ясували, що через слабку перевірку власників і вмісту зловмисник може додати до реєстру свій сервер або перехопити чужий, а метадані інструментів дають йому змогу схилити модель до потрібних дій навіть без вразливостей у коді. Для захисту автори запропонували засіб, який аналізує сервер ще до підключення [8]. Основні вектори атак на екосистему MCP і засоби протидії зведено в табл. 1.

Таблиця 1 – Вектори атак на екосистему MCP і засоби протидії

| Вектор атаки | Суть атаки | Засоби протидії |
|---|---|---|
| Отруєння опису інструмента | в описі є приховані інструкції, які бачить модель, але не бачить користувач | показ повного опису користувачеві, перевірка описів на вбудовані інструкції |
| Підміна опису після схвалення | сервер змінює опис інструмента, якому користувач уже довіряє | контрольна сума опису, повторне схвалення після зміни |
| Затінення інструментів | опис з одного сервера змінює роботу агента з інструментами іншого сервера | розмежування серверів, контроль потоків даних між ними |
| Шкідливий сервер у реєстрі | підроблений або перехоплений пакет у ланцюгу постачання | перевірка походження пакета, фіксація версій, аналіз сервера до підключення |
| Непряма ін’єкція через дані | інструкції в даних, які інструмент повертає агентові | мінімальні права агента, підтвердження чутливих дій людиною |
| Помилки реалізації | вразливості в клієнтах і проміжних засобах протоколу | вчасне оновлення, запуск серверів в ізольованому середовищі |

Як видно з розглянутих випадків, одного засобу захисту недостатньо, бо мовна модель не завжди здатна відрізнити інструкцію від даних. Тому захист варто будувати на кількох рівнях. Під час постачання адміністратор допускає лише сервери з перевіреним походженням і фіксує їхні версії, а спеціальні засоби аналізу ще до підключення перевіряють описи інструментів на вбудовані інструкції. На рівні клієнта застосунок показує користувачеві повний опис, зберігає його контрольну суму і просить повторного схвалення після кожної зміни.

Під час роботи агент має лише ті права, яких потребує поточне завдання, сервери запускаються в ізольованому середовищі, а надіслати дані назовні, змінити чи видалити ресурси агент може тільки після підтвердження людиною. Це відповідає рекомендаціям розробників протоколу: показувати користувачеві повну команду запуску локального сервера, запускати сервери з мінімальними правами і розширювати повноваження поступово [9].

## Висновки

Протокол MCP спростив підключення ШІ-агентів до зовнішніх інструментів, але водночас зробив частиною поверхні атаки текст, який читає модель. Отруєння описів інструментів спирається саме на це і не потребує класичних вразливостей, тому засоби перевірки коду його здебільшого не помічають. Ефективний захист потребує поєднання кількох рівнів: перевірки серверів до підключення, контролю незмінності описів, мінімальних прав агента та підтвердження чутливих дій людиною. Подальші дослідження варто спрямувати на автоматичне виявлення прихованих інструкцій в описах інструментів і на моделі контролю потоків даних між серверами, з якими працює агент.

## СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ


1. Introducing the Model Context Protocol / Anthropic. – 2024 [Електронний ресурс] – Режим доступу до ресурсу: https://www.anthropic.com/news/model-context-protocol
2. Model Context Protocol Specification, version 2025-11-25. Tools [Електронний ресурс] – Режим доступу до ресурсу: https://modelcontextprotocol.io/specification/2025-11-25/server/tools
3. OWASP Top 10 for Agentic Applications for 2026 / OWASP GenAI Security Project. – 2025 [Електронний ресурс] – Режим доступу до ресурсу: https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/
4. Beurer-Kellner L. MCP Security Notification: Tool Poisoning Attacks / L. Beurer-Kellner, M. Fischer // Invariant Labs Blog. – 2025 [Електронний ресурс] – Режим доступу до ресурсу: https://invariantlabs.ai/blog/mcp-security-notification-tool-poisoning-attacks
5. Milanta M. GitHub MCP Exploited: Accessing private repositories via MCP / M. Milanta, L. Beurer-Kellner // Invariant Labs Blog. – 2025 [Електронний ресурс] – Режим доступу до ресурсу: https://invariantlabs.ai/blog/mcp-github-vulnerability
6. Lakshmanan R. First Malicious MCP Server Found Stealing Emails in Rogue Postmark-MCP Package / R. Lakshmanan // The Hacker News. – 2025 [Електронний ресурс] – Режим доступу до ресурсу: https://thehackernews.com/2025/09/first-malicious-mcp-server-found.html
7. Peles O. Critical RCE Vulnerability in mcp-remote: CVE-2025-6514 Threatens LLM Clients / O. Peles // JFrog Blog. – 2025 [Електронний ресурс] – Режим доступу до ресурсу: https://jfrog.com/blog/2025-6514-critical-mcp-remote-rce-vulnerability/
8. Li X. A First Look at the Security Issues in the Model Context Protocol Ecosystem / X. Li, X. Gao // IEEE/IFIP International Conference on Dependable Systems and Networks (DSN 2026). – 2026 [Електронний ресурс] – Режим доступу до ресурсу: https://arxiv.org/abs/2510.16558
9. Model Context Protocol Specification, version 2025-11-25. Security Best Practices [Електронний ресурс] – Режим доступу до ресурсу: https://modelcontextprotocol.io/specification/2025-11-25/basic/security_best_practices



Урлапова Дар’я Павлівна — студентка бакалаврату кафедри Автоматизації та інтелектуальних інформаційних технологій, Вінницький національний технічний університет, м. Вінниця, e-mail: dashaurlapova@gmail.com.

Урлапова Марія Павлівна — студентка бакалаврату кафедри Автоматизації та інтелектуальних інформаційних технологій, Вінницький національний технічний університет, м. Вінниця, e-mail: mashaurlapova23@gmail.com.

Гармаш Володимир Володимирович — канд. техн. наук, доцент кафедри Автоматизації та інтелектуальних інформаційних технологій, Вінницький національний технічний університет, м. Вінниця, e-mail: garmash.v.v@vntu.edu.ua.

Urlapova Daria P. — undergraduate student, Faculty of intelligent information technologies and automation, Vinnytsia National Technical University, Vinnytsia, e-mail: dashaurlapova@gmail.com.

Urlapova Maria P. — undergraduate student, Faculty of intelligent information technologies and automation, Vinnytsia National Technical University, Vinnytsia, e-mail: mashaurlapova23@gmail.com.

Harmash Volodymyr V. — Cand. Sc. (Eng), Senior Lecturer, Faculty of intelligent information technologies and automation, Vinnytsia National Technical University, Vinnytsia, e-mail: garmash.v.v@vntu.edu.ua.