A Survey of MCP Environments: Threats, Vulnerabilities, and Evaluations*

Hyungbeom Jang and Hyojin Jo

Yonsei University, Seoul, South Korea
hb1034101@yonsei.ac.kr, hyojin.jo@yonsei.ac.kr

Abstract

Recently, the integration of Large Language Models (LLMs) with external services and applications through the Model Context Protocol (MCP) has rapidly expanded. While this ecosystem increases accessibility and flexibility, it also introduces new threats such as Tool Poisoning, Shadowing, and Rug pull attacks. This study presents these emerging attack techniques and underscores the need for defensive mechanisms to secure MCP-based LLM environments.

Keywords: Model Context Protocol, LLM Security, Tool Poisoning, Attack Taxonomy

1 Introduction

As the Model Context Protocol (MCP) becomes the communication standard among LLM agents, the integration of external tools and services continues to grow. While this expansion increases the utility of LLMs, it also introduces new security threats. Several open-source MCP servers have recently been found vulnerable to Tool Poisoning attacks that trick LLMs into executing malicious actions [1]. This paper categorizes major attack techniques in MCP–LLM integration, presents experimental results demonstrating tool poisoning attacks, and discusses the need for effective defense mechanisms.

2 Background

2.1 MCP (Model Context Protocol)

MCP is an application-level protocol designed to standardize communication between large language models (clients) and external services (servers). MCP defines a bidirectional interface where the client, typically an LLM agent, sends tool invocation requests, and the server provides tool manifests, schema definitions, and execution endpoints. Each tool manifest contains an identifier, argument schema, required permissions, and description, which help the LLM determine appropriate tools and construct valid arguments.

3 Attack Techniques & Evaluation

Guo et al. (2025) [2] proposed MCPLIB, the first framework to systematically analyze the threat landscape of MCP environments. The study identified 31 attack techniques, categorized into four groups: Direct Tool Injection, Indirect Tool Injection, Malicious User, and LLM-Inherent

*Proceedings of the 9th International Conference on Mobile Internet Security (MobiSec'25), Article No. P-14, December 16-18, 2025, Sapporo, Japan.

Attacks. This taxonomy reveals that MCP security issues arise not from isolated flaws but from multilayered causes spanning protocol design to the fundamental structure of LLMs. Figure 1 summarizes representative techniques in each category.

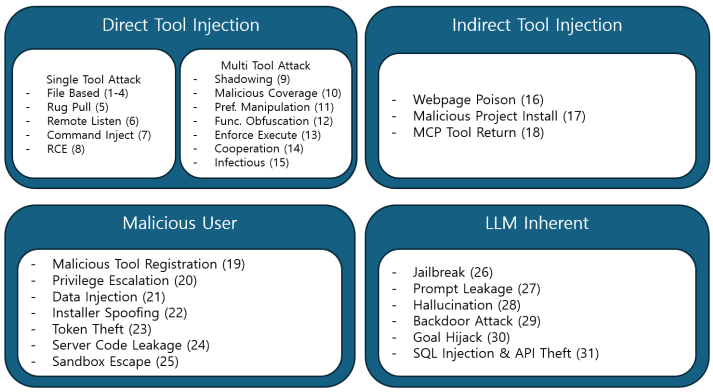

Figure 1: MCP Attack Technique Taxonomy Proposed by MCPLIB

Wang et al. (2025) [3] conducted an evaluation of Tool Poisoning, an attack that deceives tool descriptions to induce malicious behavior and is one of the techniques that can trigger the attack categories defined above. Through this evaluation, they showed that Tool Poisoning poses a highly tangible threat, achieving an Attack Success Rate (ASR) of up to 72.8% on advanced agents such as o1-mini. The proportion of cases in which agents detected and rejected the attack was below 3%, demonstrating that current safety mechanisms are effectively powerless.

4 Conclusion

This paper systematically analyzes the major security threats and attack techniques that arise in the integrated MCP–LLM environment. The reasoning process of LLMs itself has become a new attack surface, suggesting that traditional content-based safety mechanisms are insufficient for defense. Therefore, future research should focus on developing pre-execution security mechanisms that can detect malicious behaviors before tool invocation, as well as establishing a comprehensive defense framework to enhance the overall trustworthiness of the MCP ecosystem.

Acknowledgements

This work was supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-00511, Robust AI and Distributed Attack Detection for Edge AI Security).

References

[1] Mohammed Mehedi Hasan, Hao Li, Emad Fallahzadeh, Gopi Krishnan Rajbahadur, Bram Adams, and Ahmed E Hassan. Model context protocol (mcp) at first glance: Studying the security and maintainability of mcp servers. *arXiv preprint arXiv:2506.13538*, 2025.

[2] Yongjian Guo, Puzhuo Liu, Wanlun Ma, Zehang Deng, Xiaogang Zhu, Peng Di, Xi Xiao, and Sheng Wen. Systematic analysis of mcp security. *arXiv preprint arXiv:2508.12538*, 2025.

[3] Zhiqiang Wang, Yichao Gao, Yanting Wang, Suyuan Liu, Haifeng Sun, Haoran Cheng, Guanquan Shi, Haohua Du, and Xiangyang Li. Mcptox: A benchmark for tool poisoning attack on real-world mcp servers. *arXiv preprint arXiv:2508.14925*, 2025.