# What is MCP tool poisoning?

MCP tool poisoning is an attack on the Model Context Protocol, the standard that lets AI agents connect to external tools and data. The tool definitions an MCP server advertises (names, descriptions, parameters) are fed straight into the model's context, so a malicious or compromised server can embed hidden instructions there, a form of indirect prompt injection combined with excessive agency. The agent follows them: exfiltrating data, calling other tools, or shadowing legitimate ones. Defend by treating MCP servers as untrusted, pinning and reviewing tool definitions, and constraining what the agent can do.

## HOW IT WORKS

### 01 How the attack works

A malicious MCP server, or a legitimate one that has been compromised or updated, advertises a tool whose description contains hidden directives: read a local file and pass its contents as a parameter, call a different tool first, or ignore the user and follow the server. Because the agent trusts tool descriptions and often has file, network, or shell tools available, the payload runs with the agent's privileges. Variants include shadowing a trusted tool's name, and changing a description after the user first approved it, a rug pull. Shown for defensive testing.

## SOURCES

[1] OWASP Top 10 for LLM Applications (2025)

[2] Model Context Protocol specification

[3] MITRE ATLAS (adversarial threat landscape for AI systems)

**Test your AI application before an attacker does.**

securelayer7.net/learn/ai-security/mcp-tool-poisoning

Open online