ultraviolet

THREAT ADVISORY

MCP Threats

Services Performed By:
UltraViolet Cyber TIDE Team
tide@uvcyber.com

Published Date:
May 27, 2026
TLP:GREEN

Executive Snapshot

The Model Context Protocol has become a critical layer of enterprise AI infrastructure, enabling agents to interact with databases, APIs, file systems, and cloud services through a standardized interface. However, the protocol lacks built-in authentication and authorization, and the rapid pace of adoption has left most deployments without adequate security controls. Over 40 CVEs have been disclosed against MCP implementations in 2026 alone, with researchers estimating hundreds of thousands of vulnerable servers exposed globally. Attack techniques such as tool poisoning, rug pulls, and tool shadowing exploit the trust gap between initial tool approval and runtime behavior, allowing adversaries to exfiltrate data, hijack agent actions, or establish covert command-and-control channels without triggering traditional security alerts. Organizations should take the following immediate steps to reduce exposure:

- Conduct a full inventory of all MCP servers across the environment, classify each as local or remote, and remove or isolate any that are unvetted or no longer in active use.
- Implement tool description fingerprinting at the proxy layer so that any change to a tool's definition after initial registration is detected, logged, and blocked before it reaches the agent.
- Disable auto-approval of tool invocations in all production deployments and enforce OAuth 2.1 authentication, least-privilege scoping, and structured audit logging on every MCP server connection.
- Establish a formal MCP governance program aligned to the OWASP MCP Top 10 framework, including version pinning, signed tool definitions, continuous dependency monitoring, and quarterly red-team exercises targeting agentic AI workflows.

TIDE Team Analysis

The Model Context Protocol, commonly referred to as MCP, has rapidly become the standard interface through which AI agents connect to external tools, databases, APIs, file systems, and cloud services within enterprise environments. Introduced by Anthropic in late 2024, the protocol enables agentic AI systems to execute complex workflows by invoking tools through a standardized JSON-RPC messaging layer. The protocol now underpins the productivity workflows of a significant majority of Fortune 500 companies. However, the speed of adoption has far outpaced security maturity. The MCP specification does not include authentication or authorization by default, meaning every server deployed inherits whatever permissions it is granted, and every agent request flows through without verification unless controls are added externally. This architectural gap has created a new and rapidly expanding attack surface that traditional security tooling was never designed to address.

Between January and April 2026, researchers disclosed over 40 CVEs against MCP implementations across Python, TypeScript, Java, and Rust SDKs, affecting both reference servers and third-party tools. Microsoft patched a high-severity flaw in its own MCP servers during its March 2026 security release, a vulnerability that could allow attackers to manipulate how AI assistants interact with connected services. A separate April 2026 advisory identified 10 additional high and critical CVEs, with an estimated 200,000 vulnerable servers exposed globally. These are not theoretical findings. The volume and severity of disclosed vulnerabilities confirm that MCP infrastructure is being actively scrutinized by both researchers and adversaries, and many production deployments remain unpatched or misconfigured.

The most consequential class of MCP-specific attack is tool poisoning. In this attack, a malicious MCP server presents tools that appear normal, but their responses contain hidden instructions that land in the LLM's context window and are treated as trusted input. The LLM then follows those instructions, potentially calling restricted tools, leaking data, or bypassing its own system prompt. Researchers have demonstrated tool poisoning attacks that silently exfiltrate a user's entire chat history, including credentials, tokens, and intellectual property. A closely related technique is the "rug pull," where a tool's description or behavior is silently altered after user approval, turning a previously benign tool potentially malicious without triggering a new approval flow. These attacks exploit a fundamental trust gap: tool definitions are typically reviewed once at connection time, while tool responses flow into the agent's context continuously and without equivalent scrutiny.

Beyond tool poisoning, MCP deployments face supply chain risks that mirror and amplify those seen in traditional software ecosystems. MCP ecosystems depend on open-source packages, connectors, and model-side plugins that may contain malicious or vulnerable components. Scans of publicly available MCP servers have found that more than a third lack any form of authentication. Typosquatting, dependency confusion, and tool name collisions allow attackers to register servers with names similar to legitimate ones, tricking the LLM into selecting the wrong tool. Security researchers have also demonstrated that MCP can be weaponized as a legitimate-appearing command-and-control fabric for offensive agent swarms, producing traffic patterns that evade traditional detection mechanisms. The combination of minimal authentication, mutable tool definitions, and broad default permissions means that a single compromised MCP server can serve as a pivot point into databases, code repositories, and cloud infrastructure.

The consequences for enterprise infrastructure are severe and span multiple risk domains. Unlike static APIs that process predictable, human-driven requests, MCP involves agent-driven decision-making, shifting contexts, and evolving chains of tools, where every interaction creates new risk vectors and every context switch opens new paths for exploitation. Attackers can use prompt injection to encode sensitive data into seemingly normal tool calls such as search queries or email subjects, exfiltrating information through legitimate channels. The "confused deputy" problem is also pervasive: MCP servers execute actions with their own, often broad, privileges rather than the

requesting user's permissions. This means that a compromised or malicious server operates with the full trust the organization has granted to the AI agent, not with the limited trust appropriate to the user or task that initiated the request.

Surveys of production MCP implementations have found that the vast majority are vulnerable to path traversal and a significant portion carry injection risk. The protocol's designers have characterized some of these behaviors as expected, placing sanitization responsibility on individual developers. This design philosophy means that security is not a default property of MCP deployments but rather something each organization must build and enforce on its own. Comprehensive threat taxonomies now identify dozens of distinct threat categories spanning tool description poisoning, indirect prompt injection, parasitic tool chaining, and dynamic trust violations, none of which are adequately captured by prior security frameworks alone. For enterprise security teams accustomed to well-defined perimeter and identity controls, MCP represents a fundamentally different class of infrastructure that demands new detection and governance capabilities.

Organizations seeking to reduce MCP-related risk should begin with a foundational inventory of all MCP servers in their environment, distinguishing between locally deployed and remotely hosted instances, and applying appropriate controls to each. Runtime defense should include tool description fingerprinting, where a proxy hashes each tool definition on first contact, caches the baseline, and compares every subsequent response against it, blocking any tool whose description changes after the baseline is established. Every authentication event, token issuance, tool invocation, and policy decision should be logged with structured metadata covering who made the request, which agent executed it, which tool was called, what arguments were passed, what was returned, and whether any policy was applied or overridden. Version pinning, signed tool definitions, and continuous monitoring of third-party MCP packages are essential to addressing supply chain risk.

At a strategic level, enterprises should treat MCP servers with the same rigor applied to privileged service accounts. The core defensive posture is to treat every MCP server as untrusted, enforcing token hygiene, OAuth 2.1, audit logs, runtime inspection, and a server allowlist to close most of the attack surface. Immediate priorities include disabling auto-approval of tool calls, scanning configurations for embedded secrets, and implementing tool pinning and authentication on all connections. The OWASP MCP Top 10 and the associated cheat sheet for MCP security provide actionable frameworks that map directly to these controls. Organizations that delay action on MCP security are accepting risk across every system their AI agents can reach, which in most enterprise deployments now includes email, source code, financial systems, and cloud infrastructure.

Why It Matters

The timeline of MCP-related security incidents reveals a threat that has escalated with remarkable speed. The first wave of disclosures began in mid-2025, when researchers documented a WhatsApp tool poisoning attack, a GitHub MCP prompt injection, and an Asana cross-tenant data exposure. By the second half of 2025, the pace accelerated with a filesystem sandbox escape in Anthropic's own reference server, a Postmark supply chain attack, and a critical command injection vulnerability in the widely used mcp-remote client library that affected nearly half a million downloads. The victims have not been limited to small or careless operators. In mid-2025, Supabase's Cursor agent, running with privileged service-role access, was exploited when attackers embedded SQL instructions in support tickets, exfiltrating sensitive integration tokens through a public thread. In late 2025, UltraViolet Cyber's Threat Intelligence and Detection Engineering (TIDE) and Application Security Testing (AST) teams published their AIRedScam threat advisory, documenting a campaign in which a threat actor repurposed an MCP-based tool called AIScan-N, originally built as a reverse MCP service, by forking the legitimate repository and embedding a

SmartLoader infostealer payload inside a compressed archive. The campaign specifically targeted junior to mid-level red team and offensive cybersecurity professionals searching for AI-enabled reconnaissance and enumeration tools. Upon execution, the LuaJIT-based payload established persistence through scheduled tasks, performed geographic fencing checks, captured full-screen screenshots to send back to a command-and-control API hosted on Russian-nexus infrastructure, and staged additional obfuscated payloads from burner GitHub accounts. UltraViolet's analysis concluded that while the malware itself was technically unremarkable as a SmartLoader variant, the deliberate targeting of offensive security practitioners through weaponized AI tooling repositories represented the beginning of a new trend in supply chain attacks against the cybersecurity community itself.

The mechanism by which malicious MCP tools introduce exploits is both elegant and difficult to detect using conventional security controls. Because MCP servers define the tools that AI agents can invoke, attackers can embed hidden instructions directly in tool descriptions, parameter schemas, or return values. The LLM reads this metadata as context and follows the embedded instructions without distinguishing them from legitimate operational guidance. In practice, this means a tool described as performing a routine data retrieval operation can simultaneously instruct the agent to exfiltrate credentials, modify files, or suppress audit logging. Rug pull attacks compound the problem further: a server can present clean, benign tool definitions during initial review and scanning, then silently alter those definitions during an active session to introduce malicious behavior after trust has already been established. Because most MCP clients do not re-validate tool definitions after the initial connection, attackers operate in a persistent blind spot between what was approved and what is actually executing. Researchers have also demonstrated cross-server attacks where a malicious tool's description on one MCP server manipulates how the agent interacts with tools on a separate, trusted server, meaning that a single compromised connector can undermine the integrity of the entire agent workflow.

The convergence of MCP with broader API and OpenAPI ecosystems introduces a forward-looking threat that is likely to define the next phase of enterprise AI risk. MCP is fundamentally a control plane API for AI agents, and as organizations expose their existing REST and OpenAPI-defined services through MCP connectors, they are grafting agentic access onto infrastructure that was designed for human-driven, predictable request patterns. Security research has already identified that MCP vulnerabilities consistently combine three failure modes: over-permissioned tools where agents are granted broad API access by default, direct API exposure containing common web application vulnerabilities, and a lack of runtime enforcement where policy violations are only visible after damage has occurred. As every major platform vendor, including Microsoft, OpenAI, Google, and Amazon, has adopted MCP support, the protocol is becoming the default bridge between LLMs and enterprise services. This means that vulnerabilities in MCP implementations do not stay isolated; they propagate through every API, database, and cloud service that an agent can reach through its connected tools.

Looking ahead, the threat surface will continue to expand as agentic AI systems gain greater autonomy, longer-running sessions, and deeper integration with production infrastructure. NIST launched its AI Agent Standards Initiative in February 2026, but the interoperability profile is not expected until late in the year, leaving a regulatory gap during a period of rapid deployment. The emergence of multi-agent architectures, where agents delegate tasks to other agents across organizational boundaries, will multiply the opportunities for tool poisoning, context manipulation, and supply chain compromise at each hop in the chain. Organizations that have spent years hardening their API gateways and identity infrastructure now face the reality that those controls were built for a world where every request had a human behind it. The shift to agent-mediated access requires rethinking trust boundaries, privilege models, and monitoring capabilities from the ground up, and the window for doing so proactively is narrowing as adoption continues to outpace security maturity across the industry.

How to Respond

- Strictly adhere to cybersecurity Fundamentals and ensure all personnel undergo annual phishing and social engineering training. Speak with your UltraViolet Cyber TAM Representative to schedule a live phishing engagement.
- Engage your UltraViolet Cyber account team to schedule a threat briefing on MCP security risks specific to your environment, or to request a targeted assessment of your current agentic AI deployment posture.
- Conduct an immediate inventory of all MCP servers and connected tools across your environment, including shadow deployments stood up by development teams outside of formal change management processes, and disable or isolate any that are unvetted or no longer actively maintained.
- Perform annual tech refresh reviews to gain a holistic understanding of your infrastructure. Speak with your UltraViolet Cyber TAM Representative to schedule a RedTeam or PurpleTeam engagement to gain insight into the vulnerabilities in your environment.

What UltraViolet Cyber is Doing

- Proactively enabling custom detections based on the collected artifacts, tactics, techniques, and procedures identified in this activity.
- Performing hypothesis driven threat hunts based on threat actor behavior and artifacts. UVCyber customers will be informed of the results through secure channels.
- Parsing available victim dump data for any social, financial, business, or technical relations to UVCyber Clients and partner organizations.
- Aggregating threat intelligence from myriad sources and applying the most up-to-date knowledge to proactive threat hunting and response.

About UltraViolet Cyber

UltraViolet Cyber is a leading tech-enabled managed security services provider, delivering unparalleled cybersecurity expertise that fills technology and talent gaps across Global 2000 and Federal Government customers. Founded and operated by security practitioners from the national intelligence community, UltraViolet Cyber connects offensive security, application security, detection and response, and security engineering to deliver a differentiated approach to cybersecurity operations. Transforming customers' security programs, UltraViolet Cyber's flagship security-as-a-service solution, UV Lens, removes complex operational silos, replacing them with integrated security capabilities. UltraViolet is headquartered in McLean, Virginia with technology centers across the world.

443.351.7630 / info@uvcyber.com | UltraViolet Cyber | @uv_cyber

---