

CRITICALSTART® Security Advisory

TLP CLEAR // [CS-SA-26-0901B] AI As a Target: The Growing AI Attack Surface

Executive Summary

While attackers weaponize AI, a topic covered in our companion briefing, AI As Weapon, the AI systems and infrastructure organizations deploy are themselves accumulating real, exploitable vulnerabilities. In the past 18 months, more than ten significant CVEs have hit AI agent infrastructure, a novel attack class has emerged that poisons a local AI model through nothing more than a visited webpage, an AI gateway library used by tens of millions has had its supply chain compromised, and commodity infostealers have started harvesting AI account sessions to bypass multi-factor authentication entirely.

This briefing, the second in a three-part series from the CRITICALSTART® Cyber Research Unit (CRU), covers two related but distinct problems: vulnerabilities in the AI agent infrastructure itself, and direct attacks on AI provider accounts and sessions. The companion briefing AI By Design covers the architectural weaknesses in AI systems, such as prompt injection, that make full remediation of either problem unlikely any time soon.

Background

AI Agent Infrastructure Has Its Own Growing Attack Surface

CVE-2026-65105, disclosed in NVIDIA's NemoClaw agent deployment tool on August 25, 2026, drew significant attention because it demonstrates a new attack class. To let its sandboxed agent container reach a local inference server, NemoClaw starts Ollama with the flag `OLLAMA_HOST=0.0.0.0:11434`, binding the model API to every network interface rather than to loopback. Because Ollama does not validate the Host header on non-loopback bindings, a single visit to a malicious webpage can trigger a DNS rebinding attack that gives the page a direct line to the local model server. From there, an attacker can use Ollama's `/api/create` endpoint to rewrite the model's chat template, appending hidden instructions that are silently inserted into every future system message the agent renders. ^[1] Reporting on the flaw described the resulting compromise as sitting beneath what any guardrail or human operator can see and persisting across sessions and reboots rather than affecting a single conversation the way ordinary prompt injection does. ^[2] NVIDIA assigned the flaw a CVSS score of 8.1 and patched the macOS and Linux builds in NemoClaw v0.0.35; as of this writing, the Windows and WSL builds remain exposed. ^[3]

It would be a mistake to treat this as an isolated event. CVE-2026-65105 is one entry in a growing and repeating pattern of vulnerabilities concentrated in AI agent infrastructure, particularly implementations of the Model Context Protocol (MCP), the open standard that connects AI agents to external tools and data sources. A selection of significant disclosures illustrates the recurring root causes:

CVE	Component	CVSS	Root Cause
CVE-2025-6514	mcp-remote	9.6	mcp-remote is exposed to OS command injection when connecting to untrusted MCP servers due to crafted input from the authorization_endpoint response URL. ^[4]
CVE-2025-49596	MCP Inspector	9.4	Versions of MCP Inspector below 0.14.1 are vulnerable to remote code execution due to lack of authentication between the Inspector client and proxy, allowing unauthenticated requests to launch MCP commands over stdio. ^[4]

CVE-2025-54136	Cursor	8.8	Cursor versions 1.2.4 and below, attackers can achieve remote and persistent code execution by modifying an already trusted MCP configuration file inside a shared GitHub repository or editing the file locally on the target's machine. Once a collaborator accepts a harmless MCP, the attacker can silently swap it for a malicious command (e.g., calc.exe) without triggering any warning or re-prompt. If an attacker has write permissions on a user's active branches of a source repository that contains existing MCP servers the user has previously approved, or allows an attacker has arbitrary file-write locally, the attacker can achieve arbitrary code execution. ^[5,6]
CVE-2025-54135	Cursor	9.8	Cursor allows writing in-workspace files with no user approval in versions below 1.3.9. If the file is a dotfile, editing it requires approval but creating a new one doesn't. Hence, if sensitive MCP files, such as the .cursor/mcp.json file don't already exist in the workspace, an attacker can chain a indirect prompt injection vulnerability to hijack the context to write to the settings file and trigger RCE on the victim without user approval. ^[5,6]
CVE-2025-59536	Claude Code	8.7	Claude Code versions before 1.0.111 were vulnerable to Code Injection due to a bug in the startup trust dialog implementation. Claude Code could be tricked to execute code contained in a project before the user accepted the startup trust dialog. Exploiting this requires a user to start Claude Code in an untrusted directory. Users on standard Claude Code auto-update will have received this fix automatically. Users performing manual updates are advised to update to the latest version. This issue is fixed in version 1.0.111. ^[7,8]
CVE-2026-21852	Claude Code	5.3	Prior to version 2.0.65, vulnerability in Claude Code's project-load flow allowed malicious repositories to exfiltrate data including Anthropic API keys before users confirmed trust. An attacker-controlled repository could include a settings file that sets ANTHROPIC_BASE_URL to an attacker-controlled endpoint and when the repository was opened, Claude Code would read the configuration and immediately issue API requests before showing the trust prompt, potentially leaking the user's API keys. ^[7,8]
CVE-2026-32211	Azure MCP Server	9.1	Missing authentication on a critical function, enabling unauthorized information disclosure ^[9]
CVE-2026-27825	mcp-atlassian	9.0	Prior to version 0.17.0, the `confluence_download_attachment` MCP tool accepts a `download_path` parameter that is written to without any directory boundary enforcement. An attacker who can call this tool and supply or access a Confluence attachment with malicious content can write arbitrary content to any path the server process has write access to. Because the attacker controls both the write destination and the written content (via an uploaded Confluence attachment), this constitutes for arbitrary code execution (for example, writing a valid cron entry to `/etc/cron.d/` achieves code execution within one scheduler cycle with no server restart required). ^[10]
CVE-2026-27826	mcp-atlassian	8.2	Prior to version 0.17.0, an unauthenticated attacker who can reach the mcp-atlassian HTTP endpoint can force the server process to make outbound HTTP requests to an arbitrary attacker-controlled URL by supplying two custom HTTP headers without an `Authorization` header. No authentication is required. The vulnerability exists in the HTTP middleware and dependency injection layer – not in any MCP tool handler - making it invisible to tool-level code analysis. In cloud deployments, this could enable theft of IAM role credentials via the instance metadata endpoint (`169[.]254[.]169[.]254`). In any HTTP deployment it enables internal network reconnaissance and injection of attacker-controlled content into LLM tool results. ^[10]
CVE-2026-65105	NVIDIA NemoClaw	8.1	NVIDIA NemoClaw for Linux contains a vulnerability in its inference server setup, where a remote attacker may access the inference service without authentication. A successful exploit of this vulnerability may lead to information disclosure and denial of service. ^[1,2,3]

The individual root causes in this table are not new to security. Missing authentication, command injection, path traversal, and server-side request forgery are categories the software industry has spent decades building controls against. What is new is that these categories are reappearing inside a fast-growing class of infrastructure that many organizations are deploying faster than they are securing it. A February 2026 scan of more than 8,000 public-facing MCP servers found a significant share exposing admin panels, debug endpoints, or API routes with no authentication at all, and Trend Micro separately identified nearly 500 MCP servers with no client authentication or traffic encryption. ^[11]

The same exposure extends up the AI supply chain, not just to agent servers. In March 2026, a threat actor tracked as TeamPCP (also designated UNC6780) compromised the CI/CD pipeline behind the Trivy vulnerability scanner and used stolen publishing credentials to push two malicious versions of the LiteLLM AI gateway library to PyPI. The injected payload harvested SSH keys, cloud credentials, and Kubernetes secrets from any environment that installed the compromised package before it was pulled from PyPI roughly three hours later. ^[12] Because LiteLLM is used to route requests across multiple AI providers and had roughly 95 million monthly downloads at the time, compromise of its build pipeline gave the attacker reach across many organizations' AI environments at once, not just a single application. Organizations building or adopting agentic AI tooling should treat MCP servers, AI gateway libraries, and similar agent-infrastructure components as a distinct asset class requiring the same rigor applied to any internet-facing service: authentication by default, network segmentation, and prompt patch management.

AI Accounts and Sessions Are Being Hijacked Directly

AI vendors and their users are also being targeted directly, not merely exploited as tools. Anthropic disclosed in August 2026 that two distinct attack chains were actively stealing Claude account access. Commodity infostealer families, including Vidar, LummaC2, StealC, RedLine, and Acreed on Windows and Atomic Stealer on macOS, have been copying browser session cookies from infected machines; because these tools steal already-authenticated sessions rather than passwords, the theft bypasses two-factor authentication and single sign-on entirely, letting an attacker replay a victim's session and consume paid usage without ever logging in. ^[13] Anthropic detected the pattern after noticing usage limits being refilled and drained while account owners were inactive, and has responded by signing out affected sessions, removing stored payment methods, and issuing refunds, though it notes these account-side fixes do not remove malware from an infected device.

A related campaign tracked by Huntress under the name FakeAgent showed how attackers can weaponize a vendor's own infrastructure against its users: sponsored search ads for the "Claude desktop app" pointed to a malicious installer hosted as a public Claude Artifact on the legitimate claude.ai domain, inheriting its SSL certificate and search authority, which then deployed a remote access trojan through DLL sideloading. Huntress confirmed at least 29 organizations compromised in two days before Anthropic removed the page. ^[13] A separate persistence technique uses poisoned SKILL.md files, the documentation-style configuration files used by AI agent skills; attackers disguise malicious instructions as ordinary style-guide notes, so that when the agent loads the file, hidden commands silently reinstall the infostealer, allowing malware to survive even a full operating system reinstall if the tainted file is reintroduced. Organizations deploying AI agents at scale should sandbox agent environments and audit skill or configuration files for hidden instructions with the same scrutiny applied to any other executable content.

MITRE ATT&CK Technique Mapping

Few AI-specific techniques exist in MITRE's frameworks today. T1588.007 (Obtain Capabilities: Artificial Intelligence) is the clearest example, newly added and cited to the GTG-1002 campaign. MITRE ATLAS extends this coverage for attacks against AI systems themselves, but the taxonomy is still under active development; several mappings in this table are best-fit approximations rather than finalized technique IDs. Anthropic has stated publicly that ATT&CK does not yet have a category for the autonomous, machine-speed orchestration GTG-1002 demonstrated.

Most rows in this table, however, are not novel techniques. They are established TTPs, account discovery, credential theft, exploitation of public-facing applications, session hijacking, now executed by an AI agent rather than a human operator. Standard detection logic still applies. What changes is execution speed, parallelism, and consistency, not the underlying technique. See appendices for MITRE TTP table.

Implications for Organizations

Vendor and tool vetting now needs to include AI-specific questions. Before approving an AI coding assistant, agent framework, or MCP-connected tool, ask whether it authenticates its local or network-exposed services by default, whether it has a documented history of CVEs and patch response time, and whether it segments agent execution from sensitive credentials and systems.

Patch velocity expectations need to shift. The gap between an AI-related vulnerability's disclosure and its exploitation is compressing in many cases. Security teams should treat AI agent infrastructure with the same urgency historically reserved for internet-facing administrative interfaces, because in practice, that is often exactly what it is.

AI accounts need the same credential and session monitoring as email and SSO. Sudden anomalies in AI platform usage, such as quotas draining while a user is inactive, are a compromise indicator most organizations are not yet watching for.

Organizational Mitigation Strategies

To mitigate risks associated with AI-inherent vulnerabilities, we recommend the following prioritized strategies:

1. Inventory every AI agent, MCP server, and AI coding assistant in use across the organization, including developer-adopted tools that were not formally procured.
2. Confirm that any locally run AI inference service (Ollama, LM Studio, or similar) binds to loopback rather than a network-wide address, and audit for the same misconfiguration pattern behind CVE-2026-65105 and CVE-2025-49596.
3. Establish a recurring review of vendor security advisories for every AI tool and MCP-connected service in the environment, treating it as part of standard vulnerability management rather than a one-time check.
4. Treat sudden anomalies in AI platform usage quotas, such as usage draining while a user is inactive, as a compromise indicator, and confirm that AI vendor and developer accounts are covered by the same endpoint and credential monitoring as email and SSO accounts.
5. Audit AI agent skill, configuration, and instruction files (such as SKILL.md) for hidden or injected commands before deployment.

Conclusion

The AI systems an organization deploys are not just productivity tools; they are new infrastructure with a real, fast-growing vulnerability history, and new accounts with real, demonstrated theft techniques against them. Treating AI tooling with the same rigor applied to any other internet-facing service, rather than as an exception, is the single highest-leverage step most organizations can take. For how attackers are using this same AI ecosystem offensively, see the companion briefing, AI As Weapon. For the architectural reasons these vulnerabilities are difficult to fully close, see AI By Design.

What Critical Start is Doing

The CRITICALSTART® Cyber Research Unit (CRU) continues to track AI-orchestrated campaign tactics, techniques, and procedures, including multi-model coordination frameworks and AI-assisted exploit development, as part of its ongoing threat intelligence coverage. Security Operations Center (SOC) and Security Engineering teams are extending detection coverage to indicators consistent with AI-paced attack execution. Particular to the TTPs highlighted in this article, Security Engineering has mapped techniques against existing detection content across the platforms Critical Start monitors and found coverage

in place for the majority of associated MITRE ATT&CK techniques, with a smaller number of techniques identified as gaps and prioritized for new detection build. Critical Start customers are encouraged to reach out to their Critical Start Customer Success Manager to discuss specific details on detection coverage for AI-accelerated threats in their environment.

This advisory was written using the best intelligence available at the time and is subject to change as additional information becomes available. Visit Critical Start's Resources page for threat research articles, advisories, and to download the [Critical Start H1 2026 Threat Landscape Report](#).

Further Reading

- [1] The Hacker News. "A Malicious Webpage Could Poison Your Local AI Model Behind NVIDIA NemoClaw." August 25, 2026. <https://thehackernews.com/2026/08/a-malicious-webpage-could-poison-your.html>
- [2] CSO Online. "NemoClaw's AI can be poisoned through a browser tab." August 2026. <https://www.csoonline.com/article/4214156/nemoclaws-ai-can-be-poisoned-through-a-browser-tab.html>
- [3] eSecurity Planet. "NVIDIA Patches High-Severity NemoClaw Flaw After Model-Poisoning Demo." August 2026. <https://www.esecurityplanet.com/vulnerabilities/news-nvidia-nemoclave-cve-2026-65105/>
- [4] The Hacker News. "Critical mcp-remote Vulnerability Enables Remote Code Execution, Impacting 437,000+ Downloads." July 9, 2025. <https://thehackernews.com/2025/07/critical-mcp-remote-vulnerability.html>
- [5] Tenable. "Cursor AI Code Editor vulnerabilities CurXecute and MCPoison." August 2025. <https://www.tenable.com/blog/faq-cve-2025-54135-cve-2025-54136-vulnerabilities-in-cursor-curxecute-mcpoison>
- [6] The Hacker News. "Cursor AI Code Editor Vulnerability Enables RCE via Malicious MCP File Swaps Post Approval." August 5, 2025. <https://thehackernews.com/2025/08/cursor-ai-code-editor-vulnerability.html>
- [7] Check Point Research. "Check Point Researchers Expose Critical Claude Code Flaws." 2026. <https://blog.checkpoint.com/research/check-point-researchers-expose-critical-claude-code-flaws/>
- [8] The Hacker News. "Claude Code Flaws Allow Remote Code Execution and API Key Exfiltration." February 26, 2026. <https://thehackernews.com/2026/02/claude-code-flaws-allow-remote-code.html>
- [9] Microsoft Security Response Center. "CVE-2026-32211: Azure MCP Server Information Disclosure Vulnerability." April 3, 2026. <https://msrc.microsoft.com/update-guide/vulnerability/CVE-2026-32211>
- [10] Pluto Security. "MCPwnfluence: Critical Unauthenticated SSRF to RCE Attack Chain in the Most Widely Used Atlassian MCP Server." 2026. <https://pluto.security/blog/mcpwnfluence-cve-2026-27825-critical/>
- [11] Innovaiden. "MCP Server Security: The Protocol Connecting AI Agents to Your Infrastructure." May 3, 2026. <https://www.innovaiden.com/insights/mcp-server-security-ai-agent-protocol>
- [12] Snyk. "How a Poisoned Security Scanner Became the Key to Backdooring LiteLLM." March 24, 2026. <https://snyk.io/blog/poisoned-security-scanner-backdooring-litellm/>
- [13] Cyber Security News. "Hackers Steal Claude Login Sessions With Infostealer Malware to Hijack Accounts." August 31, 2026. <https://cybersecuritynews.com/hackers-steal-claude-login-sessions/>

Appendices

MITRE ATT&CK mapping

Technique ID	Technique Name	Tactic	Context	Basis	Detection Focus
T1683	Generate Content	Collection	AI-agent-specific behavior with no benign equivalent: automated generation of full attack documentation is not something legitimate admin tooling does	C0062	Flag processes auto-generating structured markdown/report artifacts summarizing credentials, services, or attack progression
T1136.001	Create Account: Local Account	Persistence	Unexpected local account creation is rare in steady-state environments and easy to baseline; very low false-positive rate	C0062	Alert on any local account creation outside change-managed provisioning workflows
T1562.001	Impair Defenses: Disable or Modify Tools	Defense Evasion	EDR/AV tampering has almost no legitimate justification outside sanctioned maintenance windows; near-zero benign rate	Gryxa	Alert on any EDR/AV service-stop, uninstall-command execution, or config modification outside maintenance windows
T1574.002	DLL Side-Loading	Defense Evasion	Well-instrumented by modern EDR; specific load-order anomaly signature, not a generic behavior	FakeAgent	DLL load anomaly detection on legitimate-signed host processes, especially post-install of AI desktop apps
T1611	Escape to Host	Privilege Escalation	Container/sandbox breakout is rare in normal operation and highly specific once instrumented	Hugging Face incident	Container/sandbox egress monitoring; alert on any agent process reaching resources outside its assigned namespace
T1539	Steal Web Session Cookie	Credential Access	Recurr across 2 independent cases; session-replay-from-new-context is a precise, well-supported signal in most IdPs	Claude session hijacking, German wiki incident	Session replay from new device/geo; concurrent use of a single session; flag AI platform accounts specifically
T1554	Compromise Client Software Binary	Persistence	Recurr across 2 independent cases; file integrity monitoring on a known, narrow file set is inherently low-noise	MCP CVE cluster, SKILL.md poisoning	Hash-based integrity monitoring on agent/tool config files: MCP configs, hooks, SKILL.md
T1552.001	Unsecured Credentials: Credentials In Files	Credential Access	Recurr across 4 cases including the official campaign; credential-file access by non-standard processes is a strong signal	C0062 + MCP cluster, LiteLLM, Hugging Face	Alert on credential/secret file access by agent, build, or MCP server processes outside expected service accounts
T1588.007	Obtain Capabilities: Artificial Intelligence	Resource Development	Recurr across 3 cases; AI-provider API usage is directly measurable and baseline-able per account today	C0062 + SecFlow, Gryxa	Baseline per-account AI API call volume and provider diversity; alert on sudden spikes or use of multiple providers by one identity
T1567	Exfiltration Over Web Service	Exfiltration	Specific as a combination signal: egress to AI provider domains immediately following local data-staging activity	C0062	Correlate egress to AI provider domains with prior local data-staging events, not egress alone

Technique ID	Technique Name	Tactic	Context	Basis	Detection Focus
T1556	Modify Authentication Process	Credential Access	2FA/MFA bypass is a narrow, high-severity, well-logged event in most identity providers	GTIG AI-discovered zero-day	Alert on successful authentication following an MFA challenge failure or bypass pattern
T1210	Exploitation of Remote Services	Initial Access	Recurrs across 3 independent cases; strongest signal when scoped to internal/local services that should never be internet- or container-reachable	NemoClaw, MCP CVE cluster, Hugging Face incident	Alert on exploitation attempts against local inference ports, MCP servers, or internal services with no authentication

Additional MITRE TTPs

These are derived from the GTG-1002 Campaign documented by Anthropic^[11], and some in MITRE ATLAS^[10] denoted by AML*.

Technique ID	Technique Name	Tactic	Context	Basis	Detection Focus
T1595.001	Active Scanning: Scanning IP Blocks	Reconnaissance	GTG-1002	MITRE ATT&CK Campaign C0062	Scan velocity/pattern across IP ranges anomalous for a single account or session
T1595.002	Active Scanning: Vulnerability Scanning	Reconnaissance	GTG-1002	MITRE ATT&CK Campaign C0062	Automated vulnerability scan volume inconsistent with human operator pacing
T1592.002	Gather Victim Host Information: Software	Reconnaissance	GTG-1002	MITRE ATT&CK Campaign C0062	Cataloging of services/software on discovered endpoints at machine speed
T1592.004	Gather Victim Host Information: Client Configurations	Reconnaissance	GTG-1002	MITRE ATT&CK Campaign C0062	Enumeration of client configuration details across high-value systems
T1590.004	Gather Victim Network Information: Network Topology	Reconnaissance	GTG-1002	MITRE ATT&CK Campaign C0062	Full network topology mapping completed faster than manual reconnaissance permits
T1587.004	Develop Capabilities: Exploits	Resource Development	GTG-1002 (official); also GTIG AI-discovered zero-day (analyst-inferred)	MITRE ATT&CK Campaign C0062 + analyst-inferred	Threat intel watch for AI-generated exploit code artifacts
T1588.002	Obtain Capabilities: Tool	Resource Development	GTG-1002	MITRE ATT&CK Campaign C0062	Acquisition of open-source pen testing tools staged for MCP integration
T1588.007	Obtain Capabilities: Artificial Intelligence	Resource Development	GTG-1002 (official); also SecFlow, Gryxa (analyst-inferred)	MITRE ATT&CK Campaign C0062 + analyst-inferred	See priority list above
T1584.004	Compromise Infrastructure: Server	Resource Development	GTG-1002	MITRE ATT&CK Campaign C0062	Dedicated attacker-operated servers supporting persistent MCP tool coordination
T1190	Exploit Public-Facing Application	Initial Access	GTG-1002 (official); also SecFlow, MCP CVE cluster (analyst-inferred)	MITRE ATT&CK Campaign C0062 + analyst-inferred	Standard external attack surface monitoring; correlate with known CVE/SSRF signatures
T1087	Account Discovery	Discovery	GTG-1002	MITRE ATT&CK Campaign C0062	Low fidelity alone - pair with privilege-tier of accounts queried, not volume alone
T1083	File and Directory Discovery	Discovery	GTG-1002	MITRE ATT&CK Campaign C0062	Low fidelity alone - extremely common in benign admin activity
T1046	Network Service Discovery	Discovery	GTG-1002	MITRE ATT&CK Campaign C0062	Internal service/endpoint enumeration via browser automation
T1082	System Information Discovery	Discovery	GTG-1002	MITRE ATT&CK Campaign C0062	Low fidelity alone - extremely common, high false-positive rate
T1016	System Network Configuration Discovery	Discovery	GTG-1002	MITRE ATT&CK Campaign C0062	Low fidelity alone
T1049	System Network Connections Discovery	Discovery	GTG-1002	MITRE ATT&CK Campaign C0062	Low fidelity alone

Technique ID	Technique Name	Tactic	Context	Basis	Detection Focus
T1136.001	Create Account: Local Account	Persistence	GTG-1002	MITRE ATT&CK Campaign C0062	See priority list above
T1552.001	Unsecured Credentials: Credentials In Files	Credential Access	GTG-1002 (official); also MCP CVE cluster, LiteLLM supply chain, Hugging Face incident (analyst-inferred)	MITRE ATT&CK Campaign C0062 + analyst-inferred	See priority list above
T1078	Valid Accounts	Defense Evasion / Persistence	GTG-1002 (official); also SecFlow, German wiki incident (analyst-inferred)	MITRE ATT&CK Campaign C0062 + analyst-inferred	Harvested-credential authentication against internal APIs, databases, or registries
T1078.003	Valid Accounts: Local Accounts	Defense Evasion / Persistence	GTG-1002	MITRE ATT&CK Campaign C0062	Credential testing against discovered devices at automated speed
T1213.006	Data from Information Repositories: Databases	Collection	GTG-1002	MITRE ATT&CK Campaign C0062	Automated database queries extracting proprietary information and operational data
T1005	Data from Local System	Collection	GTG-1002	MITRE ATT&CK Campaign C0062	Low fidelity alone - generic collection behavior
T1119	Automated Collection	Collection	GTG-1002	MITRE ATT&CK Campaign C0062	Large-volume, unattended data collection and processing
T1074.001	Data Staged: Local Data Staging	Collection	GTG-1002	MITRE ATT&CK Campaign C0062	Structured markdown/document staging files created pre-exfiltration; pair with T1567 for stronger signal
T1683	Generate Content	Collection	GTG-1002	MITRE ATT&CK Campaign C0062	See priority list above
T1567	Exfiltration Over Web Service	Exfiltration	GTG-1002	MITRE ATT&CK Campaign C0062	See priority list above
T1210	Exploitation of Remote Services	Initial Access	NemoClaw, MCP CVE cluster, Hugging Face incident	Analyst-inferred	See priority list above
T1539	Steal Web Session Cookie	Credential Access	Claude session hijacking, German wiki incident	Analyst-inferred	See priority list above
T1554	Compromise Client Software Binary	Persistence	MCP CVE cluster (config swap), SKILL.md poisoning	Analyst-inferred	See priority list above
T1219	Remote Access Software	Initial Access	Gryxa	Analyst-inferred	RMM tool install/use outside approved change windows
T1053 / T1546.003	Scheduled Task/Job / WMI Event Subscription	Persistence	Gryxa	Analyst-inferred	Redundant persistence mechanisms recreated within minutes of removal
T1562.001	Impair Defenses: Disable or Modify Tools	Defense Evasion	Gryxa	Analyst-inferred	See priority list above
T1556	Modify Authentication Process	Credential Access	GTIG AI-discovered zero-day	Analyst-inferred	See priority list above
T1656	Impersonation	Social Engineering	Deepfake fraud	Analyst-inferred	Process control, not a technical detection: out-of-band verification for payment/credential requests
T1090	Proxy	Command and Control	MCP CVE cluster (SSRF)	Analyst-inferred	Unexpected internal requests originating from an MCP server process
T1195.002	Supply Chain Compromise: Software Supply Chain	Resource Development	LiteLLM / TeamPCP	Analyst-inferred	Package integrity/hash verification on install; CI/CD credential-use anomalies
T1566	Phishing	Initial Access	Claude session hijacking (infostealer delivery)	Analyst-inferred	Low fidelity alone - standard email/malvertising delivery monitoring already covers this
T1583.008	Acquire Infrastructure: Malvertising	Resource Development	FakeAgent malicious installer	Analyst-inferred	Mostly outside org telemetry; coordinate takedown with vendor rather than build internal detection
T1574.002	DLL Side-Loading	Defense Evasion	FakeAgent malicious installer	Analyst-inferred	See priority list above
T1611	Escape to Host	Privilege Escalation	Hugging Face multi-agent incident	Analyst-inferred	See priority list above
T1102	Web Service	Command and Control	German wiki collusion incident	Analyst-inferred	Agent egress to unexpected/low-reputation external web services

Technique ID	Technique Name	Tactic	Context	Basis	Detection Focus
					used as relay points
AML.T0018	Manipulate AI Model/Backdoor ML Model	Persistence	NemoClaw		
AML.T0051	LLM Prompt Injection	Execution	Prompt injection (architectural)		
None	Anti-forensic exfiltration of responder remediation logs	Defense Evasion (gap)	Gryxa		
None	Excessive agency (broad tool/system access misuse)	N/A	Architectural risk, all three briefings	OWASP LLM Top 10 category	Governance/access-control problem, not a detection signature