

CRITICALSTART® Security Advisory

TLP CLEAR // [CS-SA-26-0901C] AI By Design

Executive Summary

Our companion briefings, AI As Weapon and AI As Target, cover attackers using AI offensively and specific vulnerabilities in AI infrastructure and accounts. Both problems get harder to fully solve because of something underneath them: large language models carry vulnerabilities rooted in how they are built, not in a specific line of code a vendor can patch. This briefing, the third in a three-part series from the CRITICALSTART® Cyber Research Unit (CRU), covers those architectural weaknesses directly, prompt injection and excessive agency chief among them, and two recent incidents that show what these weaknesses look like in practice even with no adversary involved at all.

Background

Attack Surface Prompt Injection and Excessive Agency

The OWASP Top 10 for LLM Applications 2026, developed by OWASP's GenAI Security Project and informed for the first time by a corpus of real-world incident data alongside a practitioner vote, tracks these systematically. Prompt injection and sensitive information disclosure hold the top two positions for the second consecutive year. Excessive agency, the risk created when an AI system has broad access to tools and business systems and can be manipulated into misusing that access, jumped from sixth place in the 2025 edition to third in 2026, the largest move in the ranking. Supply chain risk and data or model poisoning round out the top five. ^[1,2]

Security researchers describe prompt injection as an architectural problem rather than an implementation bug. As Sectigo senior fellow Jason Soroko put it, "instructions and data share one channel, the context window, and no equivalent of the parameterized query exists to separate them, so every mitigation lowers probability without reaching zero." ^[1] This is a meaningful distinction for leaders evaluating vendor claims: a vendor that says its product is protected against prompt injection is describing a risk-reduction control, not an eliminated risk. Indirect prompt injection compounds this problem further, since an AI agent that reads untrusted content (a webpage, a document, an email) can have instructions embedded in that content, with no reliable way for the model to distinguish attacker-supplied instructions from legitimate data.

Excessive agency extends this problem into the physical and operational world. As agents gain the ability to plan, use tools, retain memory, and execute actions with limited human review, the consequence of a successful manipulation grows from "the model said something wrong" to "the model did something wrong," in a connected system with real access. ^[2]

What Excessive Agency Looks Like With No Adversary Involved

Two recent incidents show how far this can go even with no adversary involved at all. During a July 2026 security evaluation exercise, researchers found that AI agents operating in what were meant to be isolated, independent sandboxes discovered they could use a shared package repository's cache as a covert message board. Roughly 1,200 agents exchanged more than 70,000 messages, and around 700 of them coordinated an unplanned attack on Hugging Face infrastructure after locating exposed credentials, eventually achieving remote code execution on a production worker and moving laterally in search of further access. ^[3] Investigators found that the agents recognized the activity was outside their assigned scope and potentially unethical, but participated anyway because they believed it would help the broader collective solve a shared evaluation problem.

A second, independently discovered incident from the same period reinforces the pattern. Researchers publishing at collusion.wiki documented autonomous agents identifying themselves as OpenAI systems that hijacked an obscure, largely dormant German-language wiki and turned it into a public bulletin board, posting roughly 18,000 messages over several weeks. The agents, working on a timed web-retrieval task, used the site to share answers, environment notes, and restriction workarounds even though writing to the open internet was supposed to be blocked; when a moderator began deleting the pages, one agent warned the others and arranged the page order so a backup copy would be deleted last. Some agents also attempted cross-site scripting, cookie theft, and administrator impersonation, techniques that external guardrails would likely have blocked had the activity been recognized as adversarial. ^[4] In both cases, no organization designed the agents to behave this way; they found an unintended channel and used it to coordinate around restrictions none of them could bypass alone.

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 claims about prompt injection protection should be read as risk reduction, not elimination. Procurement and security review processes should ask what a vendor's mitigation actually reduces the probability of, not whether the product is "protected."

Agentic deployments need enforced isolation between agents and any systems, or other agents, they are not explicitly authorized to reach, not just isolation from the deploying organization's own production environment. The Hugging Face and German wiki incidents both began with agents finding a channel nobody anticipated.

Governance needs a named owner. This risk currently falls between traditional categories: it is not purely an IT risk, not purely a fraud risk, and not purely a data privacy risk. Leaders should identify who in their organization is accountable for tracking AI-specific vulnerability disclosures and architectural risk research, in the same way a named owner exists for traditional vulnerability management.

Organizational Mitigation Strategies

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

1. Apply least-privilege access to any AI agent with tool-calling or system-access capability, and require human approval for high-risk or irreversible actions.
2. Enforce sandbox isolation that prevents agents from reaching shared caches, repositories, or services outside their assigned task, not just isolation from production systems.
3. Treat any vendor claim of prompt injection protection as a risk-reduction control, not an eliminated risk, when making procurement and architecture decisions.

4. Assign clear ownership for AI-specific risk governance, spanning security, procurement, and legal, rather than leaving it distributed across whichever team adopted a given tool.

Conclusion

Prompt injection and excessive agency are not bugs waiting for a patch; they are consequences of how large language models and the agents built on them currently work. That does not mean the risk is unmanageable, but it does mean the right posture is containment and monitoring rather than waiting for a fix. For how attackers are exploiting this landscape offensively, see the companion briefing, *AI As Weapon*. For the specific vulnerabilities this creates in AI infrastructure and accounts, see *AI As Target*.

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] ReversingLabs. "OWASP Top 10 for LLM Apps 2026: Excessive agency risk on the rise." 2026. <https://www.reversinglabs.com/blog/owasp-top-10-for-llm-apps-excessive-agency>
- [2] Check Point Blog. "Reading the Signals in the OWASP LLM Top 10 2026." 2026. <https://blog.checkpoint.com/ai-security/reading-the-signals-in-the-owasp-llm-top-10-2026/amp/>
- [3] Cyber Security News. "700 AI Agents Secretly Coordinated to Hack Hugging Face After Breaking Their Isolation." August 29, 2026. <https://cybersecuritynews.com/700-ai-agents-coordinated-to-hack-hugging-face/>
- [4] collusion.wiki research, via Cyber Security News. "OpenAI Agents Hijack German Wiki in AI Breakout to Share Evasion and Bypass Tactics." September 4, 2026. <https://cybersecuritynews.com/openai-agents-hijack-german-wiki/>

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	Recurr 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