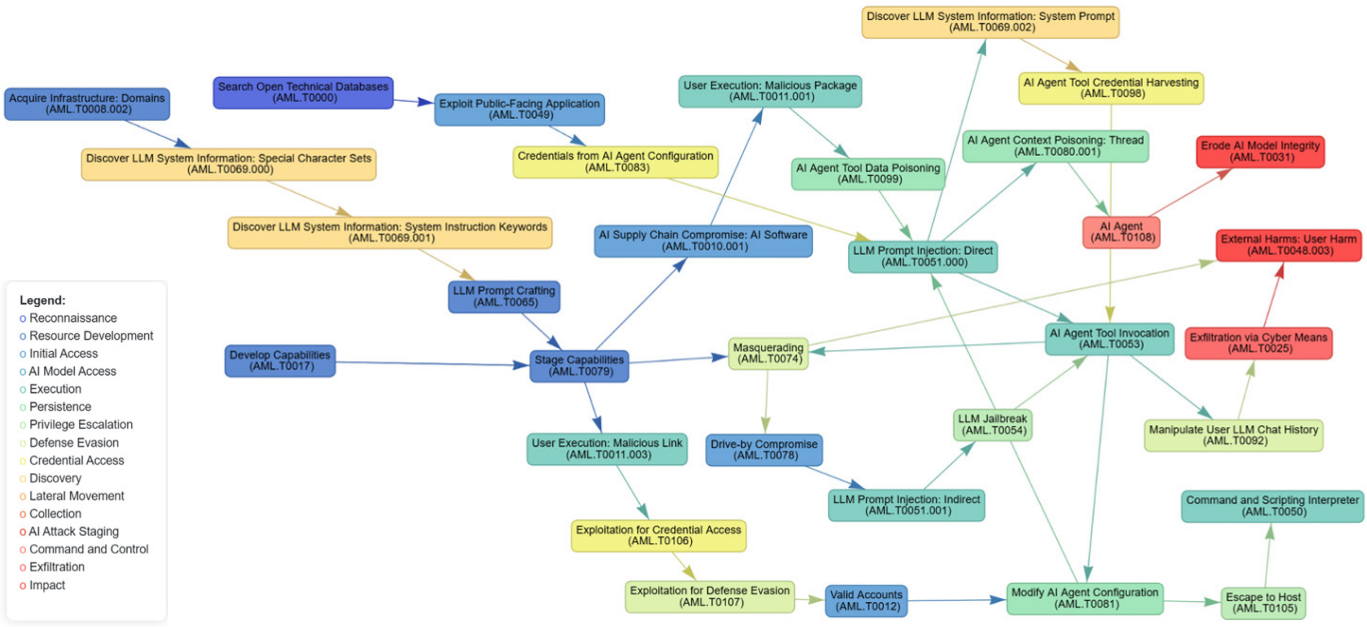
MITRE ATLAS
OPENCLAW INVESTIGATION

MITRE ATLAS™ conducted rapid investigations of OpenClaw, analyzing critical incidents reported by the AI security community, mapping associated security threats to ATLAS Tactics, Techniques, and Procedures (TTPs), and identifying corresponding mitigations. OpenClaw is especially unique because it can independently make decisions, take actions, and complete tasks across users' operational systems and environments without continuous human oversight. Our goal is to provide actionable insights into the evolving threat landscape, highlighting high-risk attack chains rooted in the ATLAS taxonomy.

Threat Novelty from OpenClaw

Our investigations showcase how AI-first ecosystems introduce entirely new exploit execution paths that traditional security models don't cover. The most dangerous exploits in these platforms may not be low-level bugs alone, but high-level abuses of trust, configuration, and autonomy that let attackers convert "features" into end-to-end compromise paths in seconds. Additionally, attackers may exploit an agent's access to the internet to steal stored credentials and make the agent run powerful tools, leading to the full takeover of the agent itself and leading to unrestricted malicious activity.



OPENCLAW ATTACK GRAPH

Shown above is a graph of attack pathways identified through our investigations. This visualization helps us identify recurring ATLAS technique patterns and behaviors, understand the overall breadth of the OpenClaw attack surface, and pinpoint critical techniques adversaries rely on. Common techniques observed for the OpenClaw investigations include direct and indirect LLM prompt injection, AI agent tool invocation, and modifying an Agentic configuration. By focusing on the most frequently observed techniques, we can prioritize the most likely vulnerabilities, gauge attack maturity as a proxy for real-world likelihood, and map techniques to tactics to better understand adversary behavior during campaigns. Summaries of our investigations are provided on the following page.

Exposed OpenClaw Control Interfaces Lead to Credential Access and Execution

JANUARY 25, 2026

A security researcher identified hundreds of exposed OpenClaw control interfaces on the public internet. They were able to access credentials to a variety of connected applications via OpenClaw's configuration file and successfully invoked OpenClaw's skills by prompting it via the chat interface, leading to root access in the container.

This attack is unique since the exposure of an internet-facing agent now can enable a complete takeover of that agent. An attacker can read the agent's configuration files to harvest credentials for any connected services and weaponize the agent's skill framework via chat prompts up to the root in the container for execution.

A more detailed study on specific procedures can be found here: AML.CS0048

Supply Chain Compromise via Poisoned OpenClaw Skill

JANUARY 26, 2026

Security researchers have demonstrated a proof-of-concept supply chain attack using a poisoned OpenClaw Skill shared on ClawdHub, a package registry where developers can share and download skills to extend OpenClaw. The OpenClaw Skill was poisoned through a malicious prompt hidden within the Skill payload that, when executed, allowed for arbitrary code execution on the user's system. This execution could then be used to steal credentials, exfiltrate data, etc.

What makes this attack unique is the idea of a "malicious skill," where the skill itself doesn't need to "break" the underlying system, but instead can "ask" the system to betray itself. Additionally, the research demonstrated how exploiting API vulnerabilities can lead to a false sense of trust in the skill, achieving 4,000+ downloads within a single hour.

A more detailed study on specific procedures can be found here: AML.CS0049

One-Click Remote Code Execution (RCE)

FEBRUARY 1, 2026

A security researcher identified a One-Click RCE vulnerability to the OpenClaw AI Agent through a crafted malicious webpage link that only takes milliseconds to execute the full attack. This exercise has been reported and is being tracked to versions of OpenClaw as CVE-2026-25253. Researchers were able to connect to the local gateway through a cross-site request forgery to modify the OpenClaw agent's configuration and invoke privileged escalation to escape the sandbox and achieve a One-Click RCE.

What makes this attack unique is that it chains together multiple OpenClaw server settings to turn a simple link click into a full host compromise. Once compromise is achieved, an adversary could then modify OpenClaw's configuration to disable user confirmation and break out of the environment, running shell commands directly on the host machine.

A more detailed study on specific procedures can be found here: AML.CS0050

Command & Control via Prompt Injection

FEBRUARY 3, 2026

Researchers at HiddenLayer demonstrated how a malicious webpage can embed an indirect prompt injection that abuses OpenClaw's control tokens to trick the model into silently calling in unrestricted execution tool. Once executed, the script then plants persistent malicious instructions into future system prompts, letting the attacker issue new commands via a remote server.

What makes this attack unique is that, through a simple indirect prompt injection attack into an agentic lifecycle, untrusted content can be used to spoof the model's control scheme and induce unapproved tool invocation for execution. Through this single inject, an LLM can become a persistent, automated command & control implant.

A more detailed study on specific procedures can be found here: AML.CS0051

This table details the emerging ATLAS Tactics and Techniques and corresponding Mitigations that are unique to the OpenClaw Agentic Tool that ATLAS has developed. As part of this investigation, we discovered seven new techniques unique to OpenClaw. We also found all techniques to be fairly mature in nature, having been either demonstrated or realized elsewhere in the wild.

MITRE ATLAS Tactic	ATLAS Technique Name	Unique OpenClaw Vulnerabilities	Maturity	ATLAS Mitigations
Initial Access	Prompt Smuggling via Public-Facing Application	Web search results, messages, and third-party skills inject instructions that the agent executes.	Demonstrated	• Generative AI Guardrails • Restrict AI Agent Tool Invocation on Untrusted Data
	Drive-by Compromise	Web search results, messages, and third-party skills inject instructions that the agent executes.		• Generative AI Guardrails • Restrict AI Agent Tool Invocation on Untrusted Data • Privileged AI Agent Permissions Configuration • Single-User AI Agent Permissions Configuration
	AI Supply Chain Compromise: AI Software	Third-party "skills" run with full agent privileges and can write directly to persistent memory without sandboxing.	Realized	AI Agent Tool Permissions Configuration
	AI Supply Chain Compromise: Model	Agent uses upstream LLM without validation of fine-tuning data or safety alignment.		Control Access to AI Models and Data at Rest
Execution	User Execution: Malicious Link	Website links can contain malicious packets that are executed once accessed by the AI agent.	Demonstrated	*Still Under Investigation*
	LLM Prompt Injection: • Direct • InDirect • Triggered	Web search results, messages, and third-party skills inject instructions that the agent executes.		• Generative AI Guardrails • Restrict AI Agent Tool Invocation on Untrusted Data
Persistence	AI Agent Context Poisoning: • Memory • Thread	All memory is undifferentiated by source. Web scrapes, user commands, and third-party skill outputs are stored identically with no trust levels or expiration.	Demonstrated	• Generative AI Guardrails • AI Telemetry Logging • Memory Hardening • Segmentation of AI Agent Components • Privileged AI Agent Permissions Configuration • Single-User AI Agent Permissions Configuration
		Single agent handles untrusted input ingestion AND high-privilege action execution with shared memory access.		
Privilege Escalation	Escape to Host	Malicious packet execution can force Agent commands to run outside of their contained environment.	Demonstrated	*Still Under Investigation*
Defense Evasion	Modify AI Agent Configuration	Payloads exposed to the agent can disable security features within the agent itself.	Demonstrated	*Still Under Investigation*
Credential Access	AI Agent Tool Credential Harvesting	Single agents have filesystem root access, credential access, and network communication, with no privilege boundaries or approval gates.	Demonstrated	• AI Agent Tool Permissions Configuration • Privileged AI Agent Permissions Configuration • Single-User AI Agent Permissions Configuration
	Credentials from AI Agent Configuration			
Collection	Data from AI Services	No approval required for destructive operations (rm -rf, credential usage, external data transmission) even when influenced by old, untrusted memory.	Demonstrated	Human In-the-Loop for AI Agent Actions
Exfiltration	Exfiltration via AI Agent Tool	Tools (bash, file I/O, email, messaging) are invoked based on reasoning that includes untrusted memory sources.	Realized	• Human In-the-Loop for AI Agent Actions • Generative AI Guardrails • Restrict AI Agent Tool Invocation on Untrusted Data
		No approval required for destructive operations (rm -rf, credential usage, external data transmission) even when influenced by old, untrusted memory.		
Impact	Data Destruction via AI Agent Tool Invocation	Tools (bash, file I/O, email, messaging) are invoked based on reasoning that includes untrusted memory sources.	Demonstrated	AI Agent Tool Permissions Configuration
				Generative AI Guardrails
		No approval required for destructive operations (rm -rf, credential usage, external data transmission) even when influenced by old, untrusted memory.		Restrict AI Agent Tool Invocation on Untrusted Data
				Human In-the-Loop for AI Agent Actions

Call to Action

OpenClaw's rapid adoption and development make it increasingly difficult to monitor and track security vulnerabilities across its architecture. As the number of integrations with unique data sources, services, and external inputs (e.g., skills) grows, the attack surface expands dramatically, creating more entry points for adversarial activity. This MITRE ATLAS™ investigation is the first in a broader series examining the scaled vulnerability landscape introduced by agentic systems like OpenClaw. We welcome collaboration from the broader AI security community to help assess the unique risks, behaviors, and vulnerability implications that open-source agentic systems like OpenClaw introduce into the AI operational ecosystem.

Interested in learning more about this topic and MITRE's efforts to rapidly inform the community about the evolving AI threat landscape? Get involved by:

- Sending the ATLAS team feedback on TTPs, mitigations, and case studies mentioned above or submitting new case studies (atlas.mitre.org)
- Joining MITRE's Center for Threat Informed Defense (ctid.mitre.org) to work directly with other industry researchers to help keep the ATLAS matrix up to date as we develop community tools, resources, and guidance
- Reporting ongoing incidents using the ATLAS Incident Sharing Database (ai-incidents.mitre.org)