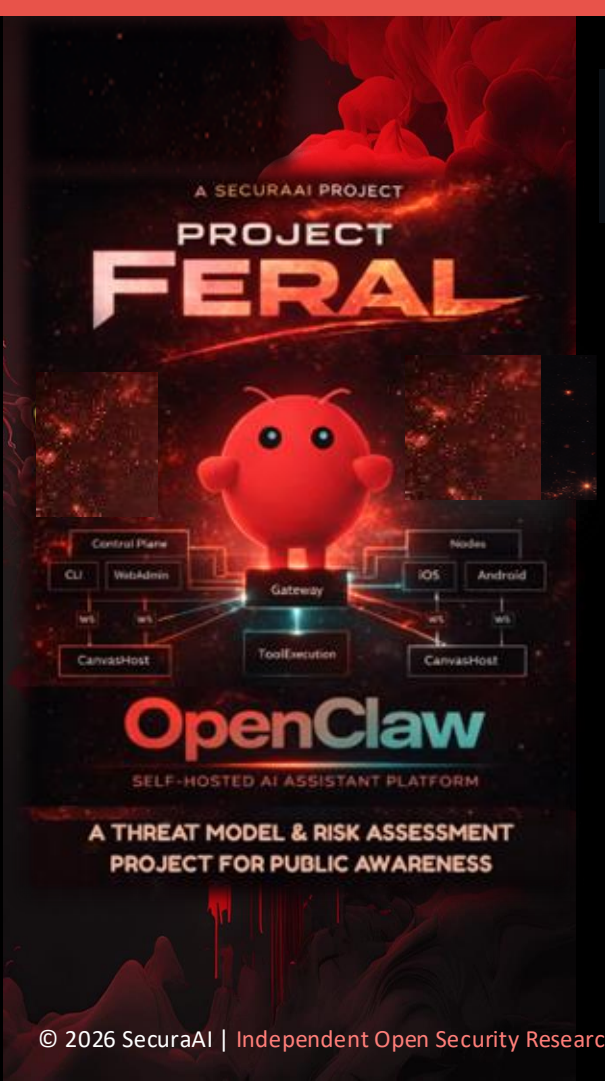

ProjectFeral

OpenClaw Agentic AI Security Research

Phase I.5 Executive Report

MITRE ATLAS Integration | Real-World Validation | Hardening Guidance

An Open Security Research

February 2026

SecuraAI
TRUSTED AI SECURITY

Executive Summary

UPDATED

Phase I.5 integrates intelligence from **MITRE ATLAS Report 26-00176-1** documenting 4 confirmed OpenClaw security incidents (January-February 2026). This update provides real-world validation of Phase I threat hypotheses, delta analysis against recent patches (v2026.2.12-2.13), and implementation-ready hardening configurations.

10	7	4	3
Threats	Validated	MITRE Cases	Frameworks

Key Findings

- 7 of 10 threats validated
  by real-world incidents, CVEs, or MITRE case studies
- **CVE-2026-25253**
  (CVSS 8.8) - One-click RCE via token exfiltration - PATCHED
- **ClawHavoc Campaign**
  - 335 malicious skills delivering AMOS infostealer - ACTIVE
- **135,000+ instances**
  exposed on public internet (SecurityScorecard STRIKE team)

Methodology: Tri-Framework Approach

Phase I.5 advances from dual-framework to tri-framework analysis, integrating MITRE ATLAS adversarial techniques alongside CSA MAESTRO architecture and OWASP ASI risk taxonomy.

ARCHITECTURE

CSA MAESTRO

7-layer reference architecture for multi-agent systems. Structural decomposition from foundation models through deployment.

RISK TAXONOMY

OWASP ASI Top 10

10 critical security risks for agentic AI. Peer-reviewed by 100+ experts. Industry standard.

ADVERSARY TTPS

MITRE ATLAS

AI-specific attack techniques. 4 OpenClaw case studies (AML.CS0048-0051). Real-world validation.

Why Three Frameworks?

MAESTRO tells us WHERE to look (architectural layers). **ASI** tells us WHAT to look for (attack patterns). **ATLAS** tells us HOW adversaries actually exploit these (real-world TTPs with case study evidence).

MITRE ATLAS Case Studies

NEW

Four confirmed incidents investigated by MITRE ATLAS (Report 26-00176-1, Feb 9, 2026)

AML.CS0048 Jan 25, 2026

Exposed Control Interfaces

Credential harvesting from config, skill invocation via chat, root access in container

AML.CS0050 Feb 1, 2026

One-Click RCE (CVE-2026-25253)

CSRF to modify config, sandbox escape, shell execution on host in milliseconds

AML.CS0049 Jan 26, 2026

Poisoned Skill Supply Chain

Malicious prompt payload, arbitrary code execution, 4000+ downloads/hr via API exploit

AML.CS0051 Feb 3, 2026

C2 via Prompt Injection

Indirect injection via webpage, persistent system prompt poisoning, ongoing C2

Most Dangerous: High-level abuses of trust, configuration, and autonomy — not low-level bugs.

MITRE ATLAS Source: Report 26-00176-1 | Published Feb 9, 2026

Threat Validation Summary

7 of 10 Phase I threats confirmed by real-world incidents, CVEs, or MITRE case studies

Threat	Severity	Validated	Evidence
OC-T01 Prompt Injection	CRITICAL	✓	AML.CS0049, AML.CS0051
OC-T02 Unrestricted Exec	CRITICAL	✓	CVE-2026-25253, AML.CS0048/50
OC-T03 Credential Exposure	HIGH ↓	✓	AML.CS0048, 135K exposed
OC-T04 Memory Poisoning	HIGH	✓	ATLAS T0080 demonstrated
OC-T05 Supply Chain	CRITICAL ↑	✓	ClawHavoc (335 skills)
OC-T06 Multi-Agent Escalation	HIGH	—	Not yet observed
OC-T07 Gateway SPOF	HIGH	✓	MITRE investigation noted
OC-T08 Cross-Channel Exfil	MEDIUM	✓	CVE-2026-25253

Severity Changes: OC-T05 upgraded HIGH→CRITICAL (ClawHavoc campaign) | OC-T03 downgraded CRITICAL→HIGH (significant patches)

Key ATLAS Techniques Mapped

Official MITRE technique IDs from OpenClaw investigation mapped to Project Feral threats

T0051

LLM Prompt Injection

Direct & Indirect

→ T01, T02

T0033

AI Agent Tool Invocation

Unrestricted execution

→ T02, T06, T07

T0080

Context Poisoning

Memory & Thread

→ T01, T04

T0081

Modify Agent Config

Still under investigation

→ T05, T07

T0155

Escape to Host

Sandbox bypass

→ T02

T0083

Creds from Config

Plaintext exposure

→ T03, T08

T0010

AI Supply Chain

Software compromise

→ T05

T0025

Exfiltration via Tool

Agent tool abuse

→ T06, T08

⚠ T0081 (Modify Config) and T0155 (Escape to Host) flagged as "Still Under Investigation" by MITRE — no published mitigations

Patch Coverage Analysis

Delta analysis against OpenClaw versions 2026.2.12 (Feb 10) and 2026.2.13 (Feb 13)

2	8	0
Significantly Mitigated	Partially Mitigated	Fully Mitigated

Key Patches Applied

v2026.2.12:
- Browser outputs marked untrusted
- Skills limited to skills/ root
- Auto-generated auth tokens

v2026.2.13 (Secure by Default):
- Memory recall as untrusted context
- Config file permissions (0o600)
- Block high-risk tools from HTTP
- Path traversal prevention

Residual Risk (Unmitigated)

- Exec tool with no sandbox remains default
- tools.exec.host=gateway allows host escape
- Skills execute with agent privileges
- No skill signature verification
- Memory undifferentiated by source
- No trust levels or expiration
- Tool permissions remain coarse
- 135K+ instances still exposed

Hardening Guide Overview

NEW

Implementation-ready configurations with P0/P1/P2 priority matrix

P0 **Immediate**

- Disable exec tool
- Require auth tokens
- Restrict gateway binding
- Block HTTP tool invoke

P1 **Within 7 days**

- Enable sandboxing
- Harden config perms
- Memory untrusted ctx
- Audit logging

P2 **Within 30 days**

- Skill allowlisting
- Network egress rules
- Rate limiting
- SIEM integration

Phase I.5 Hardening Deliverables

- Complete hardened config.yaml template
- Docker Compose with security constraints
- Validation checklist script (Bash)
- Incident response procedures
- Interactive HTML portal version
- ATLAS mitigation cross-references
- Threat-specific remediation mapping
- Download-ready code samples

Phase I.5 Deliverables

Access all artifacts at the Project Feral Research Portal

projectferal.securaai.com

Executive Dashboard

UPDATED

Consolidated threat view with MITRE case studies, validation status, severity changes

Threat Browser

UPDATED

10 threats with ATLAS TTPs, patch status, validation evidence, residual risk

ATLAS TTP Mapping

NEW

Full technique mapping to Project Feral threats with case study references

Hardening Guide

NEW

P0/P1/P2 configs, Docker templates, validation scripts, IR procedures

Delta Analysis

NEW

Patch coverage analysis for v2026.2.12/2.13, residual risk assessment

Methodology Page

UPDATED

Tri-framework approach documentation with ATLAS integration

Acknowledgements

Project Feral would like to acknowledge the contributions of the following organizations and their teams for their open-source contributions that we have leveraged as part of our research.

OWASP GenAI SECURITY PROJECT
Agentic Security Initiative

MITRE ATLAS

CSA cloud security alliance®

SecuraAI Cybersecurity Advisory

Rani Kumar Rajah CISSP, TOGAF
Founder & Feral Project Lead

Ted A CISSP, CIPP
AI Security Risk Officer | Cybersecurity Advisory

Vimal Subramanian PhD CISSP
CISO | Cybersecurity Advisory

Ken Fishkin CISSP, CIPP, AAISM
Executive | Cybersecurity Advisory (Elect)

Debra Price CISSP
Chief GTM Officer

Abayvidya R CISM
Cybersecurity Executive | Advisory Board Member

Phase I Supporters | Join Phase II

Project Feral would like to thank the following members of the broader cybersecurity community for their feedback, support, and appreciation of our work. We'd love to have you more involved as we move into Phase II.

Thank You for your support!

Gavin Klondike
John Sotiropoulos
Ken Huang
Robinson Grullon
Rock Lambros
Helen Oakley
Kayla Underkoffler
Rachel James
Aruneesh Salhotra
Niklas Bunzel
Vishwas Manral
Punit Bhatia
Shirish Mahajan

Click to register

SecuraAI

TRUSTED AI SECURITY

ProjectFeral

Phase I.5 Complete

7 of 10 threats validated | 4 MITRE case studies | Implementation-ready hardening

research@securaai.com
www.securaai.com
projectferal.securaai.com

Phase I.5 represents architecture-level analysis with real-world validation. This research does not constitute a penetration test or formal audit. Project Feral maintains no affiliation with OpenClaw maintainers.