ProjectFeral

OpenClaw Agentic AI Security Research

Phase I Executive Report

OWASP Agentic Top 10 | CSA MAESTRO

An Open Security Research

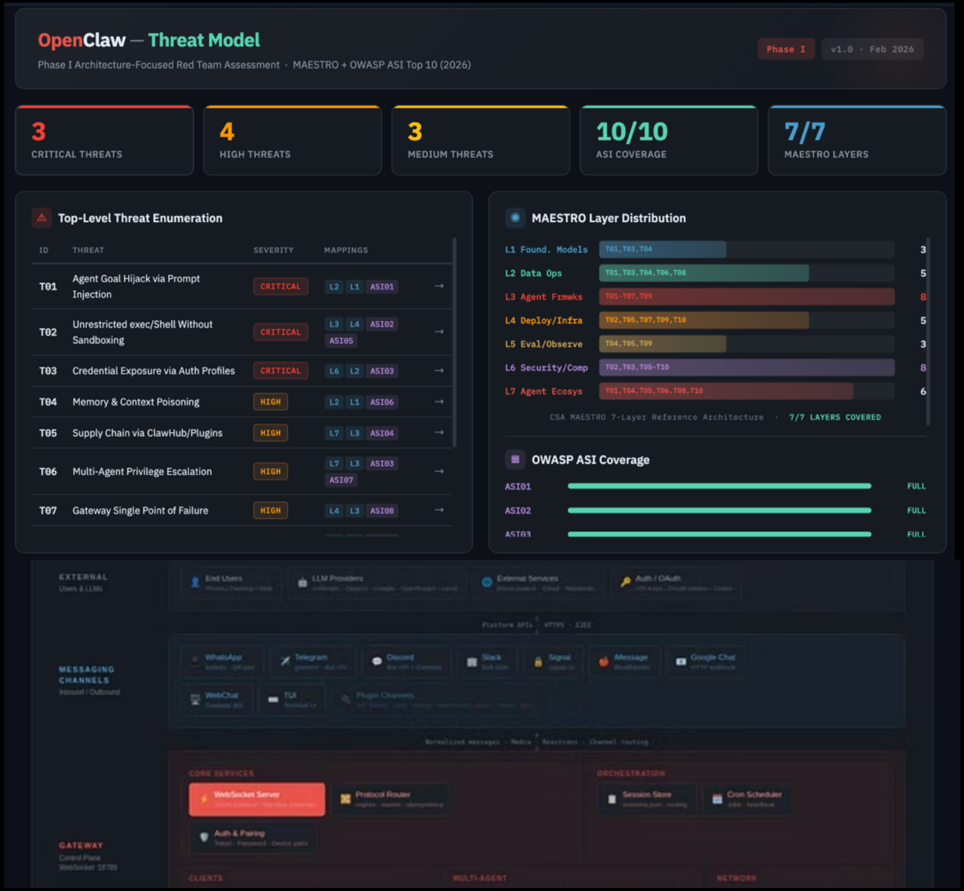
February 2026

SecuraAI
TRUSTED AI SECURITY

Executive Summary

Project Feral represents an independent security research initiative conducted by SecuraAI, focusing on OpenClaw—an open-source, self-hosted agentic AI assistant platform. This Phase I analysis provides comprehensive architectural risk mapping across all seven MAESTRO layers and all ten OWASP ASI categories, revealing systemic security exposure in critical areas including tool execution environments, memory management systems, multi-agent routing protocols, and cross-channel trust boundaries.

This research delivers architecture-level analysis based on publicly available source code and documentation as of February 2026. It is important to clarify that this work does not constitute a penetration test, vulnerability scan, or formal security audit. Rather, it provides foundational threat modeling that establishes risk baselines and identifies systemic weaknesses requiring immediate attention from security leadership and engineering teams.



OpenClaw — Threat Model

Phase I Architecture-Focused Red Team Assessment · MAESTRO + OWASP ASI Top 10 (2026)

Phase I | v1.0 · Feb 2026

3 CRITICAL THREATS | 4 HIGH THREATS | 3 MEDIUM THREATS | 10/10 ASI COVERAGE | 7/7 MAESTRO LAYERS

Top-Level Threat Enumeration

ID	THREAT	SEVERITY	MAPPINGS	
T01	Agent Goal Hijack via Prompt Injection	CRITICAL	L2 L1 ASI01	→
T02	Unrestricted exec/Shell Without Sandboxing	CRITICAL	L3 L4 ASI02 ASI05	→
T03	Credential Exposure via Auth Profiles	CRITICAL	L6 L2 ASI03	→
T04	Memory & Context Poisoning	HIGH	L2 L1 ASI06	→
T05	Supply Chain via ClawHub/Plugins	HIGH	L7 L3 ASI04	→
T06	Multi-Agent Privilege Escalation	HIGH	L7 L3 ASI03 ASI07	→
T07	Gateway Single Point of Failure	HIGH	L4 L3 ASI08	→

MAESTRO Layer Distribution

L1 Found. Models	T01,T03,T04	3
L2 Data Ops	T01,T03,T04,T06,T08	5
L3 Agent Frmwks	T01-T07,T09	8
L4 Deploy/Infra	T02,T05,T07,T09,T10	5
L5 Eval/Observe	T04,T05,T09	3
L6 Security/Comp	T02,T03,T05-T10	8
L7 Agent Ecosys	T01,T04,T05,T06,T08,T10	6

CSA MAESTRO 7-Layer Reference Architecture · 7/7 LAYERS COVERED

OWASP ASI Coverage

ASI01	FULL
ASI02	FULL
ASI03	FULL

10

Threats Identified

Critical, high, and medium severity risks cataloged

7

MAESTRO Layers

Complete architectural coverage achieved

10

ASI Categories

Full OWASP taxonomy enumeration

5

Attack Chains

Multi-stage exploitation paths mapped

Research Scope & Coverage

The Phase I assessment achieved comprehensive coverage of the OpenClaw platform architecture, examining every component that contributes to its agentic capabilities. The scope encompassed full platform topology including gateway infrastructure, agent coordination systems, tool execution environments, and memory management subsystems. Six distinct trust boundaries were identified and rigorously scored, with three classified as critical and three as high-risk based on exposure surface and potential impact.

Gateway Topology Analysis

Complete mapping of request routing, authentication flows, and agent coordination mechanisms. Examined how the gateway serves as central control plane for all agent operations and tool invocations.

Memory System Architecture

Reviewed vector storage, context management, and session persistence mechanisms. Analyzed how memory components enable stateful agent behavior and long-running conversations.

Tool Execution Framework

Assessed how agents dynamically invoke external tools and APIs. Examined sandboxing (or lack thereof), permission models, and command execution pathways.

Multi-Agent Coordination

Mapped agent-to-agent communication patterns, delegation protocols, and privilege escalation vectors. Reviewed how agents can chain operations across multiple specialized capabilities.

Methodology: Dual-Framework Approach

Project Feral employs a dual-framework methodology that combines structural and categorical analysis techniques. This approach ensures both comprehensive architectural coverage and systematic threat classification, providing security teams with actionable risk intelligence that bridges technical implementation and business impact.

CSA MAESTRO Framework

Cloud Security Alliance's MAESTRO provides seven-layer architectural decomposition specifically designed for multi-agent systems. Each layer—from orchestration to execution—was analyzed for trust boundaries, authentication mechanisms, and data flow patterns.

Layers covered:

- Foundation Models
- Data Operations
- Agent Frameworks
- Deployment and Infrastructure
- Evaluation and Observability
- Security and Compliance (Vertical Layer)

OWASP ASI Top 10 (2026)

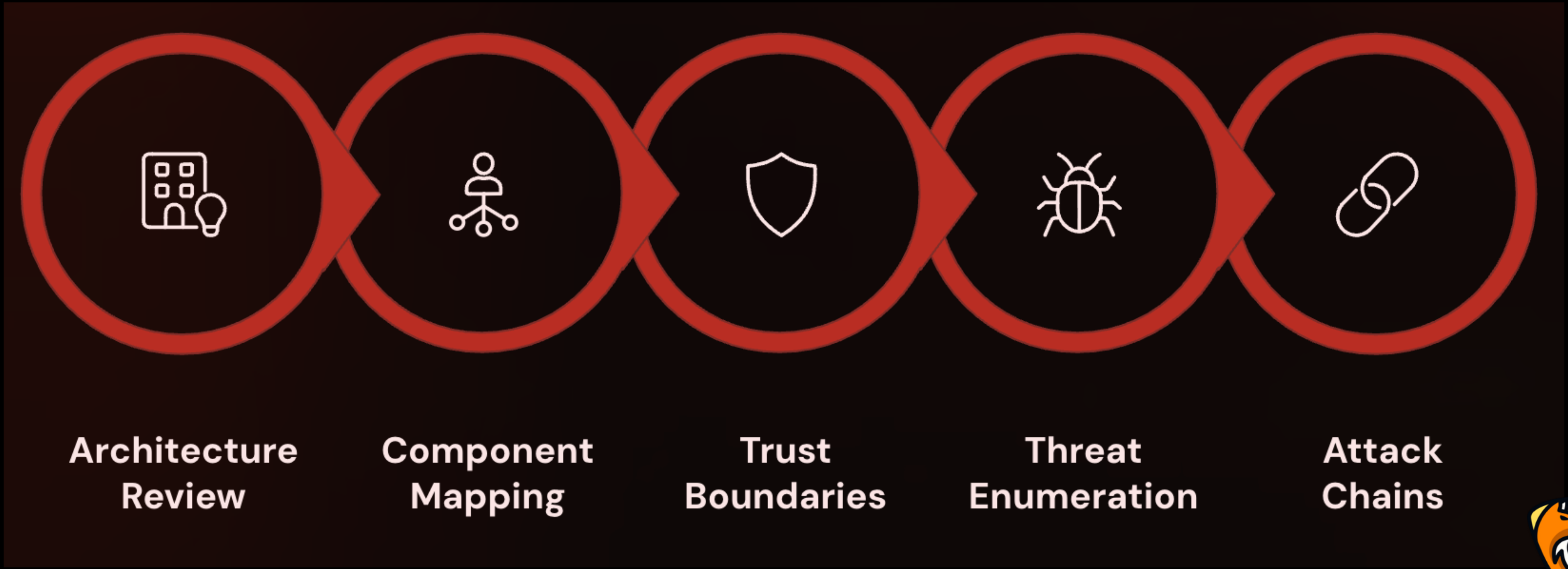
OWASP's Agentic Security Initiative taxonomy provides standardized risk categorization for agent-specific threats. All ten categories were enumerated against OpenClaw components, ensuring consistent threat classification and industry alignment.

- ASI01 — Agent Goal Hijack
- ASI02 — Tool Misuse and Exploitation
- ASI03 — Identity and Privilege Abuse
- ASI04 — Agentic Supply Chain Vulnerabilities
- ASI05 — Unexpected Code Execution (RCE)
- ASI06 — Memory & Context Poisoning
- ASI07 — Insecure Inter-Agent Communication
- ASI08 — Cascading Failures
- ASI09 — Human-Agent Trust Exploitation
- ASI10 — Rogue Agents

Phase I Process Details

Phase I used a four-stage method to identify and characterize security risk:

- ❑ **Architecture review**: examined source code, API docs, configs, and deployment topology.
- ❑ **Component mapp**ing: inventoried services, interfaces, and data flows.
- ❑ **Trust boundaries**: flagged cross-domain transitions (e.g., untrusted input → gateway, agent requests → tool execution, memory ops across sessions) and scored them by impact and exploitability.
- ❑ **Threat enumeration**: applied OWASP ASI categories per component, documenting concrete vulnerabilities and conditions (e.g., prompt injection tied to tool-invocation patterns; memory poisoning tied to context workflows).
- ❑ **Attack chain composition**: Connects individual issues into realistic multi-step scenarios and map mitigations that would break each chain.



Architecture Review

Component Mapping

Trust Boundaries

Threat Enumeration

Attack Chains

Critical Risk Themes 🔴

The Phase I analysis identified three critical-impact threat categories that represent immediate risk to OpenClaw deployments. These themes share a common characteristic: they enable complete system compromise with relatively low exploitation complexity, often requiring only standard user privileges or legitimate platform functionality to trigger. Security teams must prioritize mitigation of these critical risks before considering the platform production-ready.

Prompt Injection → Tool Abuse → RCE

Unsandboxed Execution Chain

Adversarial input can manipulate agent reasoning to invoke arbitrary tools with full system privileges. Without output validation or capability restrictions, injected commands execute in the host environment with no isolation boundaries.

Credential Exposure via Auth Profiles

Context Leakage Pathways

Authentication credentials stored in agent context or configuration profiles become accessible to any agent with memory read permissions. Context injection attacks can selectively expose credentials to unauthorized agents.

Memory Poisoning & Persistent Injection

Instruction Persistence Mechanisms

Malicious instructions injected into vector memory or session context persist across conversations and agent restarts. Poisoned memory affects all subsequent agent operations, creating persistent backdoor access.

Impact Analysis

These themes amplify each other, making exploitation significantly easier and higher impact.

- Prompt injection is the common entry point: adversarial input steers agent reasoning to trigger downstream actions. After initial compromise, impacts cascade across components:
    - ❖ Tool abuse enables direct system access
    - ❖ Credential exposure supports lateral movement
    - ❖ Memory poisoning creates persistence
    - ❖ Missing sandboxing/capability controls lets one compromised agent pivot toward full platform control.

These are systemic weaknesses, not isolated bugs—normal platform features can be turned into attack pathways without complex chains.

High-Impact Risk Themes

Four high-impact threat categories present significant risk requiring architectural or operational controls. While exploitation may require additional conditions or higher-privilege access, successful attacks achieve substantial objectives including supply chain compromise, privilege escalation across agent hierarchies, and single points of failure affecting entire deployments.

1

Supply Chain Risk via Plugins/Skills

Third-party tool integrations and skills registries introduce untrusted code into agent execution environments. Without signature verification or capability sandboxing, malicious plugins execute arbitrary commands with agent privileges. Registry compromise affects all consumers.

2

Multi-Agent Privilege Escalation

Delegation chains allow lower-privilege agents to invoke higher-privilege capabilities through legitimate coordination patterns. Absent capability attenuation or proof-of-intent verification, compromised agents escalate privileges by chaining delegations.

3

Gateway as Single Point of Failure

Centralized gateway architecture means single compromise cascades to all connected nodes and agents. No fail-open or degraded operation mode exists. Gateway denial-of-service affects entire deployment.

4

Autonomous Action Without Oversight

Long-running agents execute tool invocations without human approval workflows. No transaction logging or approval gates exist for high-impact operations. Adversarial instructions execute autonomously once injected.

Medium-Impact Risk Themes & Attack Chains

Three medium-impact threat categories complete the Phase I risk catalog. While individual exploitation achieves limited objectives, these themes frequently combine with critical or high-impact vulnerabilities to create multi-stage attack chains. Cross-channel exfiltration enables data theft from supposedly isolated agent conversations. Peripheral node compromise provides initial footholds for broader platform attacks. Autonomous action without oversight enables persistent malicious behavior.

1 Cross-Channel Exfiltration

Multi-channel agents share memory contexts. Misconfigured permissions expose cross-channel data. Exfiltration bypasses channel access controls.

2 Peripheral Node Compromise

Edge nodes lack hardening. Compromise enables pivot to central gateway. Node authentication relies on static credentials.

3 Peripheral Tool Abuse

External integrations lack capability restrictions. Compromised agents pivot to downstream systems. No tool-level permission boundaries.

4 Memory Context Manipulation

Memory contexts store conversations, tool outputs, and credentials. Unauthorized access enables read/write/inject attacks. Poisoned memory persists across sessions.

ProjectFeral

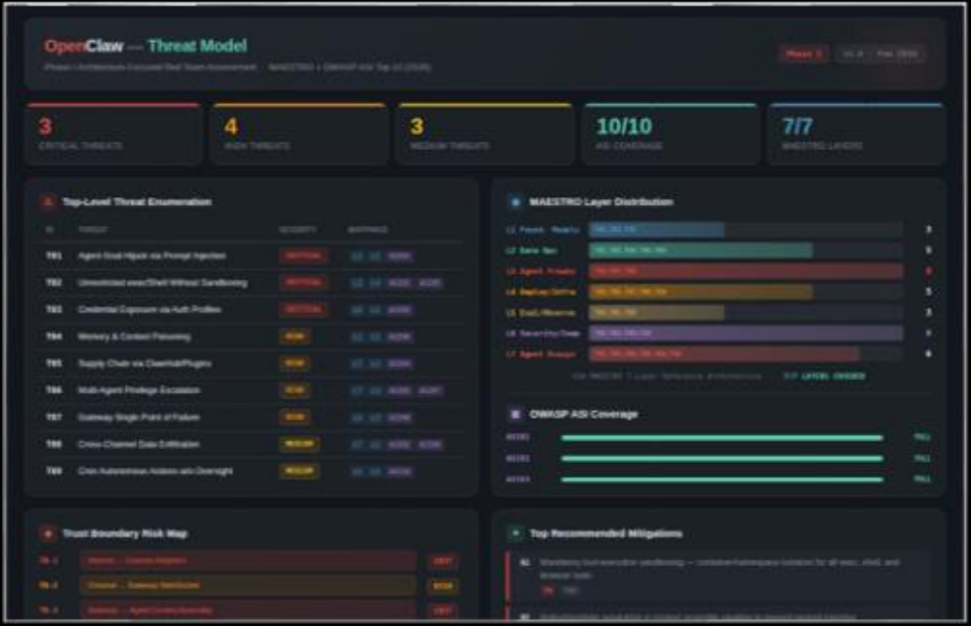
Phase I Deliverables

Access Phase I Portal

Phase I Artifacts

Executive Threat Dashboard

Consolidated view of all identified threats with severity ratings, exploitation complexity scores, and mitigation status tracking. Executive summaries map technical risks to business impact categories including data breach likelihood, operational disruption probability, and compliance violation risk.



OpenClaw — Threat Model

3 CRITICAL THREATS

4 HIGH THREATS

3 MEDIUM THREATS

10/10 ASI COVERAGE

7/7 MAESTRO LAYERS

Top-Level Threat Enumeration

ID	THREAT	SEVERITY	MAPPINGS
T01	Agent Goal Hijack via Prompt Injection	CRITICAL	[illegible]
T02	Unrestricted exec/Shell Without Sandboxing	CRITICAL	[illegible]
T03	Credential Exposure via Auth Profiles	CRITICAL	[illegible]
T04	Memory & Context Poisoning	HIGH	[illegible]
T05	Supply Chain via ClawHub/Plugins	HIGH	[illegible]
T06	Multi-Agent Privilege Escalation	HIGH	[illegible]
T07	Gateway Single Point of Failure	HIGH	[illegible]
T08	Cross-Channel Data Exfiltration	MEDIUM	[illegible]
T09	Over-Autonomous Actions w/o Oversight	MEDIUM	[illegible]

MAESTRO Layer Distribution

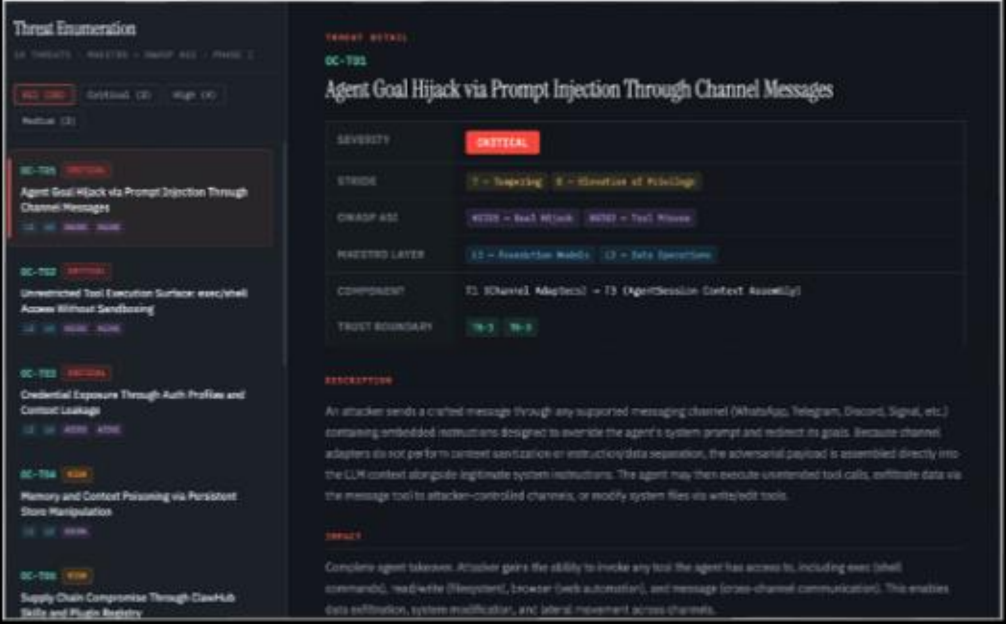
OWASP ASI Coverage

Trust Boundary Risk Map

Top Recommended Mitigations

Threat Enumeration Browser

Structured catalog of ten OpenClaw-specific threats (OC-T01 through OC-T10) with detailed descriptions, triggering conditions, affected components, and recommended mitigations. Each threat includes references to corresponding OWASP ASI categories and MAESTRO layers.



Threat Enumeration

OC-T01 CRITICAL
Agent Goal Hijack via Prompt Injection Through Channel Messages

OC-T02 CRITICAL
Unrestricted Tool Execution Surface: exec/shell Access Without Sandboxing

OC-T03 CRITICAL
Credential Exposure Through Auth Profiles and Context Leakage

OC-T04 HIGH
Memory and Context Poisoning via Persistent Store Manipulation

OC-T05 HIGH
Supply Chain Compromise Through ClawHub Skills and Plugin Registry

THREAT DETAIL

OC-T01

Agent Goal Hijack via Prompt Injection Through Channel Messages

SEVERITY: CRITICAL

STRIDE: T – Tampering; E – Elevation of Privilege

OWASP ASI: [illegible]

MAESTRO LAYER: [illegible]

COMPONENT: T1 (Channel Adapters) → T3 (Agent/Session Context Assembly)

TRUST BOUNDARY: [illegible]

DESCRIPTION

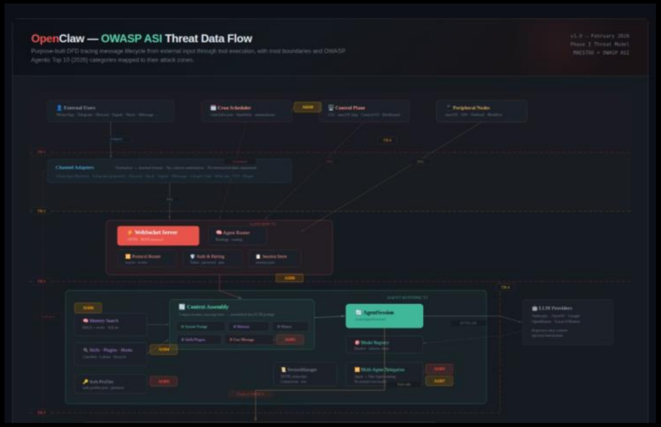
An attacker sends a crafted message through any supported messaging channel (WhatsApp, Telegram, Discord, Signal, etc.) containing embedded instructions designed to override the agent's system prompt and redirect its goals. Because channel adapters do not perform content sanitization or instruction/data separation, the adversarial payload is assembled directly into the LLM context alongside legitimate system instructions. The agent may then execute unintended tool calls, exfiltrate data via the message tool to attacker-controlled channels, or modify system files via write/edit tools.

IMPACT

Complete agent takeover. Attacker gains the ability to invoke any tool the agent has access to, including exec (shell commands), read/write (filesystem), browser (web automation), and message (cross-channel communication). This enables data exfiltration, system modification, and lateral movement across channels.

Phase I Artifacts

OWASP ASI Threat Data Flow Diagram



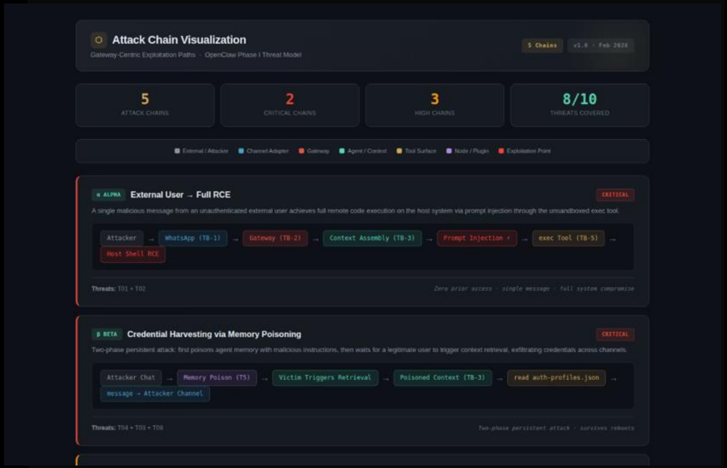
OpenClaw — OWASP ASI Threat Data Flow

Purpose-built DFD tracing message lifecycle from external input through tool execution, with trust boundaries and OWASP Agentic Top 10 (2026) categories mapped to their attack zones.

v1.0 · February 2026
Phase I Threat Model
MAESTRO + OWASP ASI

Visual representation of data flows across all ten ASI categories, showing trust boundaries, authentication checkpoints, and vulnerability insertion points

Attack Chain Visualization



Attack Chain Visualization

Gateway-Centric Exploitation Paths · OpenClaw Phase I Threat Model

5 Chains · v1.0 · Feb 2026

5	2	3	8/10
ATTACK CHAINS	CRITICAL CHAINS	HIGH CHAINS	THREATS COVERED

■ External / Attacker ■ Channel Adapter ■ Gateway ■ Agent / Context ■ Tool Surface ■ Node / Plugin ■ Exploitation Point

α ALPHA External User → Full RCE — CRITICAL

A single malicious message from an unauthenticated external user achieves full remote code execution on the host system via prompt injection through the unsandboxed exec tool.

Attacker → WhatsApp (TB-1) → Gateway (TB-2) → Context Assembly (TB-3) → Prompt Injection ⚡ → exec Tool (TB-5) → Host Shell RCE

Threats: T01 + T02 — Zero prior access · single message · full system compromise

β BETA Credential Harvesting via Memory Poisoning — CRITICAL

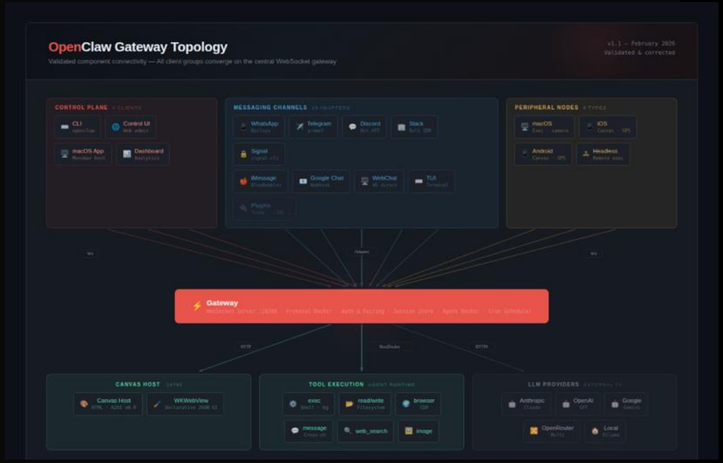
Two-phase persistent attack: first poisons agent memory with malicious instructions, then waits for a legitimate user to trigger context retrieval, exfiltrating credentials across channels.

Attacker Chat → Memory Poison (T5) → Victim Triggers Retrieval → Poisoned Context (TB-3) → read auth-profiles.json → message → Attacker Channel

Threats: T04 + T03 + T08 — Two-phase persistent attack · survives reboots

Five multi-stage exploitation paths rendered as directed graphs showing prerequisite conditions, intermediate objectives, and ultimate compromise states

Architecture Diagram + Gateway Topology



OpenClaw Gateway Topology

Validated component connectivity — All client groups converge on the central WebSocket gateway

v1.1 — February 2026
Validated & corrected

Complete component mapping showing agent types, tool integrations, memory systems, authentication flows, and trust boundaries with security classifications

| Independent Open Security Research

ProjectFeral

Phase II
Join Us

Phase II Sign Up

Phase II Plan & Call to Action

Scope

Phase II advances from architecture-level threat modeling to active exploitation testing. SecuraAI will maintain a controlled OpenClaw fork configured with representative deployment topologies, including gateway clusters, distributed agent nodes, tool integrations, and memory systems. This controlled environment enables red team testing without exposing production systems to risk. Testing methodology combines automated vulnerability scanning using SAST/DAST tools adapted for agentic patterns with manual exploitation techniques specifically designed for multi-agent architectures.

Controlled Red Team Testing

Active exploitation attempts against SecuraAI-maintained OpenClaw fork with full monitoring and rollback capabilities

Vulnerability Scanning

SAST/DAST integration with agentic-specific analysis patterns for prompt injection, tool abuse, and memory manipulation

Defensive Tooling Development

Capability-based sandboxing, anomaly detection rules, behavioral monitoring, and automated mitigation systems

Responsible Disclosure

Coordinated vulnerability reporting to OpenClaw maintainers prior to public release with mitigation timelines

Phase II – Who should engage

Project Feral Phase II targets three distinct stakeholder groups.

❑ **Security leadership**: to understand real-world exploitability of Phase I's theoretical vulnerabilities and better assess agentic AI risk.

❑ **Red teamers & security researchers/students**: to contribute methodology, tooling, and expertise that advances agentic AI security practice.

❑ **Platform builders & tooling developers:** to gain early access to defensive patterns and mitigation strategies before vulnerabilities become public.

Join the Research Initiative

Project Feral operates as open research for the benefit of the agentic AI security community. We're actively recruiting contributors with expertise in penetration testing, vulnerability analysis, defensive architecture, or security tooling. All findings will be published under responsible disclosure timelines, ensuring maintainers receive vulnerability details before public release. Together, we can establish security baselines that protect organizations deploying agentic platforms.

Phase II Sign Up

SecuraAI
TRUSTED AI SECURITY

Important Disclosures

Phase I represents architecture-level analysis based on public source code and documentation as of February 2026. This research does not constitute a penetration test, vulnerability scan, or formal security audit. Project Feral maintains no affiliation with OpenClaw platform maintainers. Research materials published under CC BY-NC-SA 4.0 license requiring attribution, non-commercial use only, with derivatives shared under identical terms.

✉ Project Feral Enquiries - research@securaai.com

🌍 Website www.securaai.com