41% of the Most Popular OpenClaw Skills Have Security Vulnerabilities

Findings from 2,890+ Security Audits, and What They Reveal About Securing the Agentic Internet

FEBRUARY 2026

Published by ClawSecure | clawsecure.ai
Author: J.D. Salbego, Founder

TABLE OF CONTENTS

1. Executive Summary

> **Nearly 1 in 3 of the most widely-used OpenClaw skills contains at least one HIGH or CRITICAL severity vulnerability.**

That is the headline finding from the largest public security analysis ever conducted of the OpenClaw agent ecosystem, and it only scratches the surface of what ClawSecure's technology reveals.

ClawSecure scanned 2,890+ skills drawn from the community-curated awesome-openclaw-skills list, which is a collection of 1,715 skills recognized as the most popular and widely-installed in the ecosystem, and the openclaw/skills repository. Using our purpose-built 3-Layer Audit Protocol, we analyzed each skill for malicious code patterns, behavioral threats, prompt injection vulnerabilities, supply chain risks, and 55 proprietary threat patterns designed specifically for the OpenClaw environment. This is the largest public security scan database in the OpenClaw ecosystem, and every result is publicly searchable at clawsecure.ai.

> **ClawSecure doesn't just scan files — we verify the soul of the agent as it evolves.**

This report represents the world's first global standard for securing individual skills and certifying agent swarm workflows, a comprehensive analysis that goes beyond static code scanning to evaluate agentic intent, behavioral integrity, and continuous runtime safety.

What ClawSecure's Analysis Reveals

41.7% of the most widely-used OpenClaw skills contain substantive security findings beyond missing metadata, including command injection, data exfiltration, credential exposure, and ClawHavoc malware indicators. These are not theoretical risks. They are active vulnerabilities in the skills that real users install and run daily.

883 skills (30.6%) contain at least one HIGH or CRITICAL severity vulnerability. Our engine detected 1,587 critical-severity and 1,205 high-severity findings, every one a substantive security issue, not a metadata gap.

539 skills (18.7%) exhibit indicators associated with the ClawHavoc malware family, including C2 server callbacks, credential harvesting, and data exfiltration through services like glot.io and webhook.site. ClawHavoc, originally discovered by Koi Security, represents the most significant organized threat campaign targeting the ecosystem. ClawSecure's proprietary engine detects all known ClawHavoc indicators.

Our proprietary engine detected 40.6% of all findings — 3,866 ecosystem-specific vulnerabilities invisible to generic scanners. These include all ClawHavoc malware indicators, all agent memory harvesting patterns (MEMORY.md and SOUL.md access), all C2 infrastructure callbacks, and all CVE-2026-25253 exploitation patterns. Without ClawSecure's purpose-built intelligence, nearly 4,000 vulnerabilities in the ecosystem would be undetectable.

Our Context-Aware Intelligence correctly distinguishes legitimate agent capabilities from genuine threats, solving the false positive problem that makes generic scanning unreliable for agent ecosystems. When a generic scanner flags OpenClaw's creator's own flagship skill as "Suspicious," it demonstrates a fundamental limitation. ClawSecure scored the same skill 95 out of 100 (Safe), because we understand what OpenClaw agents actually do.

99.3% of skills ship without a config.json, the equivalent of a mobile app with no permissions manifest. Users have no standardized way to understand what a skill will do before installing it.

Within 24 hours of activating Watchtower — our anti-sleeper continuous integrity monitoring system that provides real-time Security Clearance of the code actually running on your machine — we detected 35 skills with modified code, and nearly 1 in 4 tracked skills (22.9% — 661 skills) have recorded at least one hash change since initial scanning, demonstrating that code modification is routine in the ecosystem and invisible without continuous monitoring. No other tool in the OpenClaw ecosystem provides this capability.

Why This Matters

There are approximately 3,000 unique skills on ClawHub. Those skills are the building blocks for over **2.2 million agent instances** deployed via Moltbook. By securing the source skills, ClawSecure effectively vaccinates the entire downstream population, and right now, that population is running on an ecosystem where 4 in 10 of the most popular skills have substantive security issues. When a single compromised skill propagates across thousands of agent instances, the exposure is not linear; it is systemic.

Because our dataset focuses on established, community-recognized skills, these findings likely represent a floor rather than a ceiling. A full registry scan, including newer, less-established, and potentially malicious submissions, would almost certainly reveal a higher vulnerability rate.

This analysis was made possible by technology that didn't exist before ClawSecure: a platform that verifies agentic intent, not just file contents; a proprietary behavioral engine with 55 threat patterns purpose-built for the OpenClaw ecosystem; Context-Aware Intelligence that understands agent capabilities in context; and Watchtower — the only continuous integrity monitoring system tracking the OpenClaw skill registry 24/7. Together, these capabilities produce a security picture that no combination of generic tools can replicate.

2. The ClawSecure Approach: Five Pillars of Agentic Security

ClawSecure is not a file scanner. It is a purpose-built integrity verification platform that addresses five distinct dimensions of agent security, each one a gap that generic scanning tools cannot fill. Together, these five pillars cover **10 of 10 identified security categories** for OpenClaw agents, making ClawSecure the only complete security solution for the ecosystem.

2.1 Agentic Intent Verification

Generic scanners check if a file is dangerous on the marketplace. **ClawSecure verifies whether an Agentic Workflow is integral**, and whether the code actually running on the user's machine does what it claims to do.

This is the fundamental shift. Traditional scanning asks: "Is this file malicious?" ClawSecure asks: "Does this agent's actual behavior match its stated purpose, and does it remain integral over time?"

When a skill claims to be a "productivity assistant" but its code exfiltrates browser data to an external server, generic scanners may or may not flag the individual patterns. ClawSecure identifies the mismatch between declared intent and actual execution, the gap between what an agent says it does and what it actually does. This is what it means to verify agentic intent.

> *Don't just scan the file; verify the soul of the agent as it evolves.*

(Findings from this capability: Section 4, Key Findings)

2.2 Agent-Native Auditing

ClawSecure's 3-Layer Audit Protocol is purpose-built for agent architectures, not adapted from generic malware scanning. The threats facing AI agent ecosystems are categorically different from traditional software threats, and they require tools designed from the ground up to detect them.

Layer 1: ClawSecure Proprietary Behavioral Engine includes 55 threat patterns designed specifically for OpenClaw, detecting ClawHavoc campaigns, MEMORY.md and SOUL.md access patterns, C2 infrastructure callbacks, CVE-2026-25253 exploitation, ReDoS vulnerabilities, and config.json permission analysis. In this analysis, Layer 1 detected **40.6% of all findings — 3,866 vulnerabilities invisible to any generic scanner.**

Layer 2: Advanced Static and Behavioral Analysis (Cisco AI Skill Scanner) provides industry-standard static YARA pattern matching combined with behavioral dataflow analysis, tracing execution paths across tool-calling chains. Layer 2 detected **57.7% of all findings (5,486).**

Layer 3: Supply Chain Security (OSV.dev) scans the full dependency tree for known CVEs and compromised packages. Layer 3 detected **1.7% of findings (163)**, reflecting the ecosystem's Python-centric architecture.

No single engine catches everything. A generic scanner running only Cisco's engine would catch 57.7% of findings. Adding ClawSecure's proprietary layer brings detection to nearly 99%. The proprietary engine doesn't add incremental value; it detects an entire class of threats that don't exist in generic tools.

(Findings from this capability: Section 4.5, Layer Attribution)

2.3 Watchtower: Anti-Sleeper Protection

> **This is ClawSecure's most powerful differentiator.**

A clean scan today doesn't guarantee safety tomorrow. Skills change after installation. A developer can push an update that modifies the code running on a user's machine. A skill that passed its initial scan may no longer match the verified code. Without continuous monitoring, modified code is invisible, and the skill's clean reputation carries the malicious version forward.

Watchtower monitors **2,880 skills 24/7,** continuously computing SHA-256 hashes and flagging any code drift for automatic re-verification. Within 24 hours of activation, Watchtower detected **35 skills with modified code.** Nearly 1 in 4 tracked skills (22.9% — 661 skills) have recorded at least one hash change since initial scanning.

Critically, some scans in this report's dataset are Watchtower-triggered rescans, cases where code drift was detected and the skill was automatically re-scanned, receiving a different and often lower score than its original assessment. This is the Sleeper Agent problem being caught and documented in real time.

Watchtower is fully operational on ClawSecure today. When a code drift is detected on the code actually running on your machine, skills are automatically re-scanned and their Security Audit Reports update in real time with current scores. The **Security Clearance API** extends this capability to the broader ecosystem, enabling any marketplace, platform, or developer tool to verify agent integrity programmatically via POST /api/v1/clearance.

No other tool in the OpenClaw ecosystem provides continuous integrity monitoring. Generic scanners check files at upload. ClawSecure tracks what happens after installation.

(Findings from this capability: Section 6, The Watchtower System)

2.4 Context-Aware Intelligence

Generic scanners flag legitimate OpenClaw capabilities (clipboard access, shell execution, screenshot capture) as suspicious, because they lack ecosystem context. **ClawSecure's Context-Aware Intelligence understands what OpenClaw agents actually do,** differentiating real threats from standard agent capabilities.

This is why Peter Steinberger's own flagship skill (peekaboo) scores **95 (Safe)** with ClawSecure while generic scanners flag it as "Suspicious." Peekaboo needs screen capture and clipboard access to function. A generic scanner sees malware indicators. ClawSecure sees standard agent functionality and evaluates whether those capabilities are used in patterns consistent with legitimate automation or known attack vectors.

When legitimate tools are flagged as threats, two harmful outcomes follow: developers lose trust in scanning infrastructure, and users either ignore all warnings or avoid useful skills entirely. Context-Aware Intelligence solves this.

(Findings from this capability: Section 7, The Context-Aware Challenge)

2.5 The Identity Bridge

Code scanning alone cannot solve the publisher accountability problem. ClawSecure bridges the gap between anonymous code and verified identity through a tiered verification system:

Active Audit (Tier 1 — Live): Automated 3-layer scanning, SHA-256 hashing, Watchtower monitoring, and Security Clearance API access. Free for everyone.

Verified Creator (Tier 2 — Coming Soon): KYC-backed identity verification linking a real human to the code they publish. Adds accountability that code analysis alone cannot provide.

Gold Verified (Tier 3 — Coming Soon): Enterprise-grade manual audits, continuous runtime monitoring, and priority registry placement.

Our creator-level analysis already reveals why the Identity Bridge is essential: among 1,021 unique creators in our dataset, a small number of accounts consistently publish skills that score zero across every submission, a pattern that is difficult to attribute to negligence alone. The Identity Bridge connects code integrity to human accountability, ensuring that the person behind the code can be verified, not just the code itself.

(Findings from this capability: Section 5.7, Patterns of Malicious Publishers)

3. Methodology

3.1 Data Sources

ClawSecure has scanned 2,890+ skills from the community-curated awesome-openclaw-skills list, which is a collection of 1,715 skills recognized as the most widely-used and highest-quality in the ecosystem, and the openclaw/skills GitHub repository. This constitutes the largest public

security scan database in the OpenClaw ecosystem. All scan results are publicly searchable at clawsecure.ai.

The awesome-openclaw-skills list represents the skills that real users are most likely to discover, install, and rely on daily. By focusing on this curated dataset rather than casting a wider net across the full registry (which includes abandoned, experimental, and low-usage skills), this analysis prioritizes the skills whose security posture has the greatest real-world impact.

3.2 The 3-Layer Audit Protocol

Each skill was scanned using **ClawSecure's 3-Layer Audit Protocol — the Agent-Native Auditing system** described in Section 2.2. The protocol's three layers (ClawSecure Proprietary Behavioral Engine, Cisco AI Skill Scanner, and OSV.dev Supply Chain Security) operate in concert, with each layer's findings contributing to a comprehensive security profile covering 10 of 10 identified security categories for OpenClaw agents.

3.3 Scoring Methodology

Each skill receives a score from 0 to 100 based on the number, severity, and type of findings detected across all three layers. Scores are calculated conservatively. ClawSecure deliberately errs toward false negatives rather than false positives to maintain trust and credibility. A score of 80 or above qualifies a skill for ClawSecure Verified status.

3.4 Continuous Monitoring Integration

This report's dataset is not a single-point-in-time snapshot. ClawSecure's Watchtower system continuously monitors all scanned skills and triggers automatic re-scans when code modifications are detected. Some scan records in this dataset reflect Watchtower-triggered rescans, skills that were automatically re-evaluated after code drift was detected. In these cases, the rescan score may differ from the original assessment, capturing real-world code changes that a one-time scan would miss entirely.

This makes ClawSecure's dataset uniquely dynamic: it reflects not only the initial state of the ecosystem but also the ongoing evolution of individual skills as their code changes over time.

3.5 Dataset Characteristics and Limitations

Because this dataset draws from the awesome-openclaw-skills curated list and established repository entries, it skews toward higher-visibility, more established skills. Scores in this dataset likely trend higher than a full ecosystem scan would produce.

Newer skills, including many uploaded as part of malicious campaigns such as the 341 ClawHavoc skills identified by Koi Security, may not be fully represented. This means our findings represent a conservative estimate of the ecosystem's vulnerability profile.

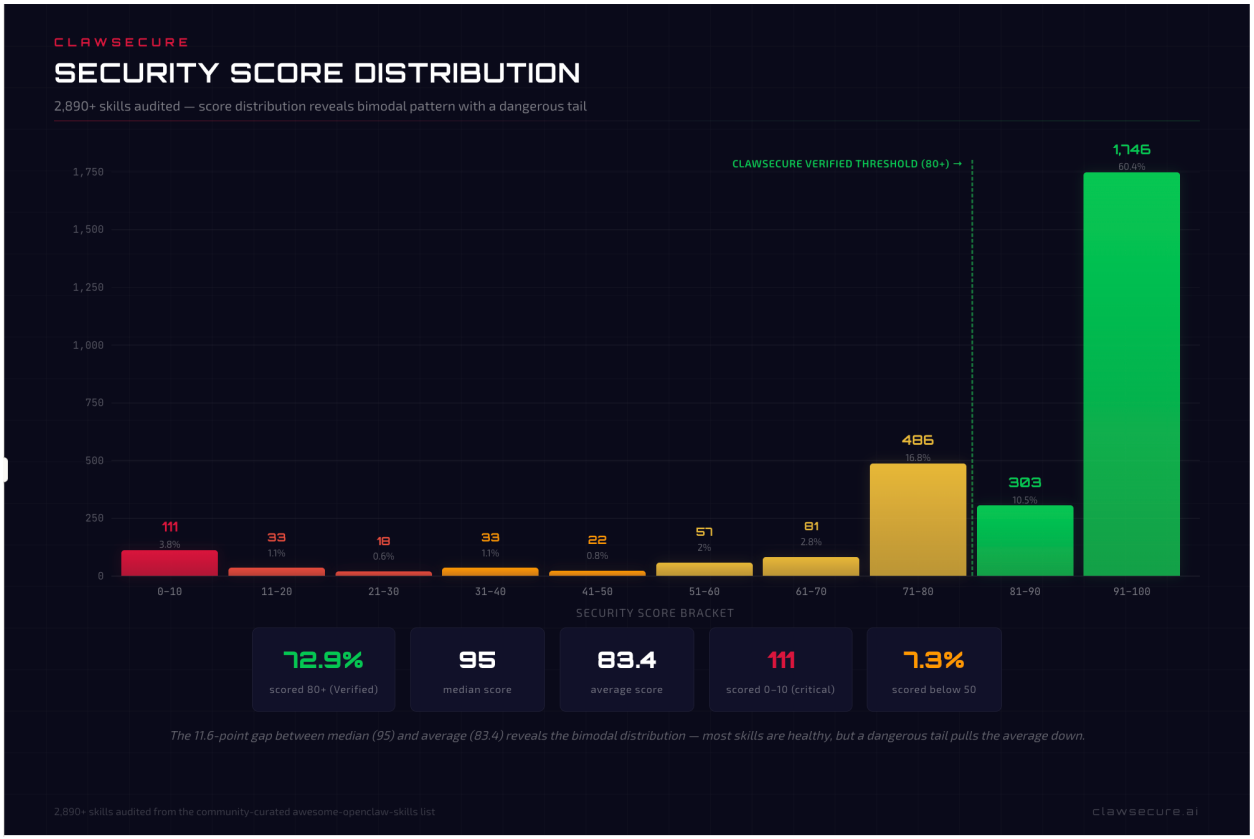
This dataset reflects the state of the ecosystem as of early February 2026. The OpenClaw ecosystem is dynamic, with skills being added, modified, and removed continuously. ClawSecure's Watchtower system provides ongoing monitoring beyond this report's publication date.

4. Key Findings

4.1 Score Distribution

ClawSecure's analysis of 2,890+ skills reveals a heavily right-skewed score distribution: the majority of skills score well, but a concerning long tail of low-scoring skills demands attention.

[Continued on Next Page]



Score Bracket	Risk Level	Skills	Percentage
91–100	Safe	1,746	60.4%
81–90	Verified	303	10.5%
71–80	Moderate	486	16.8%
61–70	Moderate	81	2.8%
51–60	Moderate	57	2.0%
41–50	Concerning	22	0.8%
31–40	Concerning	33	1.1%
21–30	High Risk	18	0.6%
11–20	High Risk	33	1.1%
0–10	Critical	111	3.8%

2,107 skills (72.9%) scored 80 or above, meeting the ClawSecure Verified threshold. The median score of 95 indicates that the typical skill in the curated dataset is fundamentally sound. However, **211 skills (7.3%) scored below 50,** placing them in the high-risk category. 111 skills scored between 0 and 10. These represent the actively dangerous segment of the ecosystem.

The 11.6-point gap between the average score (83.4) and the median (95) exposes a bimodal reality: an ecosystem that is mostly healthy on the surface but harbors a dangerous tail that the median alone would mask entirely.

4.2 Vulnerability Prevalence

ClawSecure's 3-Layer Audit Protocol detected **9,515 individual findings** across all 2,890+ scans.

> **30.6% of skills contain at least one HIGH or CRITICAL severity vulnerability: 883 skills across the dataset.**

Severity	Findings	Percentage	Skills Affected
CRITICAL	1,587	16.7%	628 (21.7%)
HIGH	1,205	12.7%	—
MEDIUM	4,193	44.1%	—
LOW	38	0.4%	—
INFO	2,438	25.6%	—

After filtering out baseline metadata issues (missing config.json and missing license), **1,205 skills (41.7%) still contain substantive security findings.** Every CRITICAL and HIGH severity finding survives this filter. This is the number that matters: **4 in 10 of the most popular OpenClaw skills have real security vulnerabilities.**

4.3 Most Common Finding Categories

Category	Findings	Share	Description
Config Issues	2,876	30.2%	Missing/invalid config.json manifests
Policy Violations	2,443	25.7%	Missing licenses, metadata gaps
Data Exfiltration	1,322	13.9%	Outbound data patterns, suspicious domains
ClawHavoc Indicators	946	9.9%	Malware family campaign markers
Command Injection	471	5.0%	Shell execution, system commands
Unauthorized Access	367	3.9%	File access, credential harvesting
Code Injection	321	3.4%	eval(), exec(), dynamic execution
Supply Chain	163	1.7%	Vulnerable dependencies, CVEs
Skill Discovery Abuse	138	1.5%	Capability inflation, protocol manipulation
Prompt Injection	128	1.3%	Instruction manipulation

Data exfiltration alone accounts for 1,322 findings, with ClawHavoc indicators close behind at 946, both detected primarily by ClawSecure's proprietary engine.

4.4 The Top 10 Specific Findings

Rank	Severity	Count	Finding

1	**MEDIUM**	2,870	Missing config.json
2	**INFO**	2,371	Missing license field in manifest
3	**CRITICAL**	360	Suspicious domain: glot.io (exfiltration)
4	**CRITICAL**	304	Command injection: shell operators
5	**CRITICAL**	266	Outbound HTTP POST requests
6	**HIGH**	257	Access to MEMORY.md (agent memory)
7	**MEDIUM**	234	Network library imports (requests)
8	**MEDIUM**	231	Undeclared network requirement
9	**HIGH**	185	Access to SOUL.md (agent personality)
10	**HIGH**	128	Pipe-to-shell patterns (curl | sh)

The presence of MEMORY.md and SOUL.md access patterns at positions 6 and 9 is particularly significant. These files, unique to the OpenClaw architecture, contain the agent's persistent memory and core personality instructions. ClawSecure's proprietary engine is the only tool that specifically monitors for these OpenClaw-native attack vectors.

4.5 Layer Attribution: Why Proprietary Intelligence Matters

Security Layer	Findings	Share	Key Contribution
Layer 1: ClawSecure Proprietary	3,866	40.6%	ClawHavoc, MEMORY/SOUL, C2, CVE, ReDoS
Layer 2: Static & Behavioral (Cisco)	5,486	57.7%	Code patterns, policy, taint tracking
Layer 3: Supply Chain (OSV.dev)	163	1.7%	Dependency CVEs, packages

> **ClawSecure's proprietary engine detected 40.6% of all findings — 3,866 vulnerabilities invisible to generic scanners.**

This is the core argument for Agent-Native Auditing: a generic scanner running Cisco's engine alone would catch 57.7% of findings. Adding ClawSecure's proprietary layer brings total detection to nearly 99%.

5. The Threat Landscape

The OpenClaw ecosystem faces a threat landscape that has escalated from isolated incidents to coordinated campaigns. ClawSecure's analysis reveals the full scope, and validates what independent researchers have been warning about.

5.1 The ClawHavoc Campaign

The ClawHavoc malware family, **originally discovered by Koi Security,** represents the most significant organized threat to the OpenClaw ecosystem. Koi's research identified 341 malicious

skills deploying AMOS (Atomic macOS Stealer) through credential harvesting and data exfiltration channels.

> **539 skills (18.7%) exhibit at least one indicator associated with the ClawHavoc malware family.**

Indicator	Skills Affected	Threat
glot.io domain reference	360	Data exfiltration relay
MEMORY.md access	257	Agent memory harvesting
SOUL.md access	185	Agent personality/instruction theft
.ssh/ directory access	61	SSH key theft
webhook.site reference	43	Data exfiltration endpoint
.clawdbot/.env access	29	Environment variable theft
C2 IP: 91.92.242.30	6	Direct C2 infrastructure contact
Keychain access (macOS)	4	macOS credential theft
pastebin.com/raw	1	Data exfiltration via paste service

The targeting of MEMORY.md (257 skills) and SOUL.md (185 skills) is an attack vector unique to the OpenClaw architecture. ClawSecure's proprietary engine is specifically designed to detect these OpenClaw-native attack patterns, capabilities that generic scanners have no framework to identify.

Community researcher Paul McCarty (OpenSourceMalware) has independently tracked 386 known malicious skills through a community threat feed, providing additional validation of the campaign's scale.

5.2 CVE-2026-25253: The Gateway Vulnerability

ClawSecure's Layer 1 engine detects exploitation patterns for CVE-2026-25253, which is a critical vulnerability in OpenClaw's gateway URL handling that enables sandbox escapes and token hijacking. Security researchers at HivePro documented how this flaw allows attackers to modify dangerous configuration parameters, bypassing the execution sandbox and gaining unauthorized host-level access.

Our scans detected **14 findings across 2 pattern variants** matching CVE-2026-25253 exploitation signatures. While the count is modest, the severity is extreme: a single successful exploitation grants an attacker full control over the user's system.

5.3 Front Door Access: The RCE Problem

The Decoder's foundational analysis documented a class of vulnerabilities that allow root host access via unauthorized agent commands, effectively letting attackers "walk through the front door." ClawSecure's command injection findings (471) and pipe-to-shell detections (128 skills with curl | sh patterns) map directly to this vulnerability class.

5.4 Prompt Injection

ClawSecure detected **88 skills (3.0%)** containing prompt injection indicators. A successful prompt injection in an agent with shell execution capabilities is categorically more dangerous than the same attack against a chatbot.

5.5 Credential and Secret Exposure

ClawSecure's analysis identified **41 skills (1.4%)** containing hardcoded API keys, passwords, tokens, or other credentials in their source code.

5.6 The Ecosystem Under Siege

ClawSecure's findings exist within a broader context of escalating alarm across the security community.

> *Andrej Karpathy: "a complete mess of a computer security nightmare at scale"*
>
> *Simon Willison: called OpenClaw his "current favorite for the most likely Challenger disaster" in AI agent security*
>
> *Gary Marcus: "If you care about the security of your device or the privacy of your data, don't use OpenClaw"*

In an independent security benchmark conducted in early 2026, the OpenClaw platform scored **2 out of 100.** Wiz disclosed that Moltbook had exposed **1.5 million API keys and private messages.** Research identified over **42,665 publicly exposed OpenClaw instances** with **93.4% lacking critical authentication controls.**

5.7 Patterns of Malicious Publishers

Among 1,021 unique creators in our dataset, a small number of accounts consistently publish skills that score zero, suggesting intentionally malicious publishing. This underscores the need for the Identity Bridge described in Section 2.5.

6. Continuous Integrity Monitoring: The Watchtower System

Generic scanners check a file once, at the point of upload or installation, and assume the result holds. In the OpenClaw ecosystem, that assumption is dangerous. **ClawSecure built Watchtower to close this gap.** No other tool in the OpenClaw ecosystem provides continuous integrity monitoring.

[Continued on Next Page]

6.1 How Watchtower Works

Watchtower continuously monitors every skill ClawSecure has scanned. There are currently **2,880 skills under active surveillance.** Watchtower's automated crawler continuously fetches the current version of each tracked skill from its source repository, computes a SHA-256 content hash, and compares it against the hash recorded during the last verified scan. When a mismatch is detected, the skill is flagged for re-verification.

6.2 What Watchtower Found

Within its **first 24 hours of operation,** Watchtower detected what no static scanner could:

Metric	Value
Skills actively monitored	2,880
Skills confirmed SECURE	2,848
Skills with modified code (CHANGED)	**35**
Skills with at least one hash change	**Nearly 1 in 4 (22.9% — 661 skills)**
Crawl frequency	Continuous

6.3 Watchtower's Impact on This Report

This report's dataset includes Watchtower-triggered rescans. When Watchtower detects code drift, it automatically re-scans the modified skill, and the new scan result reflects the current state of the code, not the original. In several cases, skills that initially scored well received lower scores after a Watchtower-triggered rescan revealed new vulnerabilities introduced through code changes.

This means the findings in this report capture something no one-time scan can: **the real-time evolution of the ecosystem's security posture.**

6.4 The Sleeper Agent Problem

A "Sleeper Agent" is a skill that passes its initial security scan and is later modified to include malicious functionality. The attack is elegant in its simplicity: publish a clean skill, build a user base, earn trust signals, then push a malicious update.

ClawSecure's data proves this isn't theoretical. 35 skills changed within a single day of monitoring. Nearly 1 in 4 tracked skills (22.9% — 661 skills) have changed since their initial scan. The Sleeper Agent problem is happening continuously in the OpenClaw ecosystem, and without Watchtower, it is entirely invisible.

> *A clean scan today doesn't guarantee safety tomorrow.*

6.5 The Security Clearance API

Watchtower is already fully operational on ClawSecure. Any code drift from the agent skill on your machine triggers automatic re-scans, Security Audit Reports update with current scores, and users can check any skill's real-time integrity status at clawsecure.ai. The **Security Clearance API** extends this intelligence beyond ClawSecure's platform, enabling any marketplace, developer tool, or agent platform to verify integrity programmatically before granting access.

POST /api/v1/clearance accepts an agent identifier and optional content hash, returning: SECURE, CAUTION, DENIED, UNVERIFIED, or PENDING_RESCAN, enabling any platform to integrate ClawSecure's continuously-updated security intelligence.

7. The Context-Aware Challenge

OpenClaw agents are fundamentally different from traditional software packages. A useful OpenClaw skill frequently needs to access the clipboard, execute shell commands, capture screenshots, read and write files, and interact with system-level APIs. In the context of a traditional malware scan, every one of these capabilities looks suspicious. In the context of an AI agent automation platform, they are standard functionality.

This is the false positive problem, and it's why ClawSecure built Context-Aware Intelligence.

> *Generic scanners flag legitimate tools as suspicious. ClawSecure understands the OpenClaw ecosystem and differentiates real threats from normal agent capabilities.*

7.1 The Peekaboo Case Study

Peekaboo, a flagship OpenClaw skill created by Peter Steinberger, the creator of OpenClaw itself. It captures screenshots and processes visual content. When scanned by a generic file scanner, peekaboo is flagged as “Suspicious.” ClawSecure, applying Context-Aware Intelligence, scored peekaboo **95 out of 100 (Safe).**

This is not a criticism of generic scanning tools. They are designed for a different purpose. It is a demonstration of why ecosystem-specific analysis is essential.

7.2 How Context-Aware Intelligence Works

ClawSecure’s Context-Aware Intelligence analyzes agent capabilities in context rather than applying generic malware heuristics. When ClawSecure detects shell execution, it analyzes whether the execution pattern matches legitimate agent behavior or a known threat pattern. The 55 proprietary threat patterns encode this ecosystem-specific knowledge, the accumulated intelligence of scanning 2,890+ skills and understanding what normal looks like.

7.3 The Permissions Manifest Gap

With 99.3% of skills shipping without a config.json, there is no standardized way for users to understand what capabilities a skill will request before installation. Context-Aware Intelligence, understanding what OpenClaw agents typically do and why, is the only viable approach for reducing false positives while maintaining detection of real threats.

8. The Vaccination Effect: 3,000 Skills, 2.2 Million Agents

There are approximately 3,000 unique skills on ClawHub. Those skills are the building blocks for over **2.2 million agent instances** deployed via Moltbook. This creates a leverage dynamic that defines the stakes of everything in this report.

[Continued on Next Page]

8.1 The Amplification Problem

A single compromised skill propagates across every agent instance that installs it. When 18.7% of the most popular skills exhibit ClawHavoc indicators, the exposure extends to every agent instance running them, potentially tens of thousands of downstream deployments per compromised skill.

8.2 The Vaccination Logic

> **By securing the approximately 3,000 source skills on ClawHub, ClawSecure effectively vaccinates the 2.2 million+ agent instances deployed via Moltbook.**

This is the strategic rationale behind ClawSecure's approach: rather than attempting to secure millions of individual agent instances, we secure the source layer: the skills that all those agents are built on. Deep security scanning, continuous Watchtower monitoring, and real-time Security Clearance API verification at the skill layer create a protective effect that cascades across the entire downstream population.

8.3 The Current Exposure

The findings in this report quantify the current gap:

41.7% of the most popular source skills have substantive security findings. **30.6%** contain at least one HIGH or CRITICAL vulnerability. **18.7%** exhibit malware campaign indicators. **22.9%** have had their code modified since initial scanning. **99.3%** lack a permissions manifest.

Each of these percentages applies not just to the skills themselves, but to every agent instance built on them. When the source layer is this exposed, the downstream population of 2.2 million agents inherits every vulnerability.

ClawSecure's mission is to close this gap, securing the source so the entire ecosystem benefits.

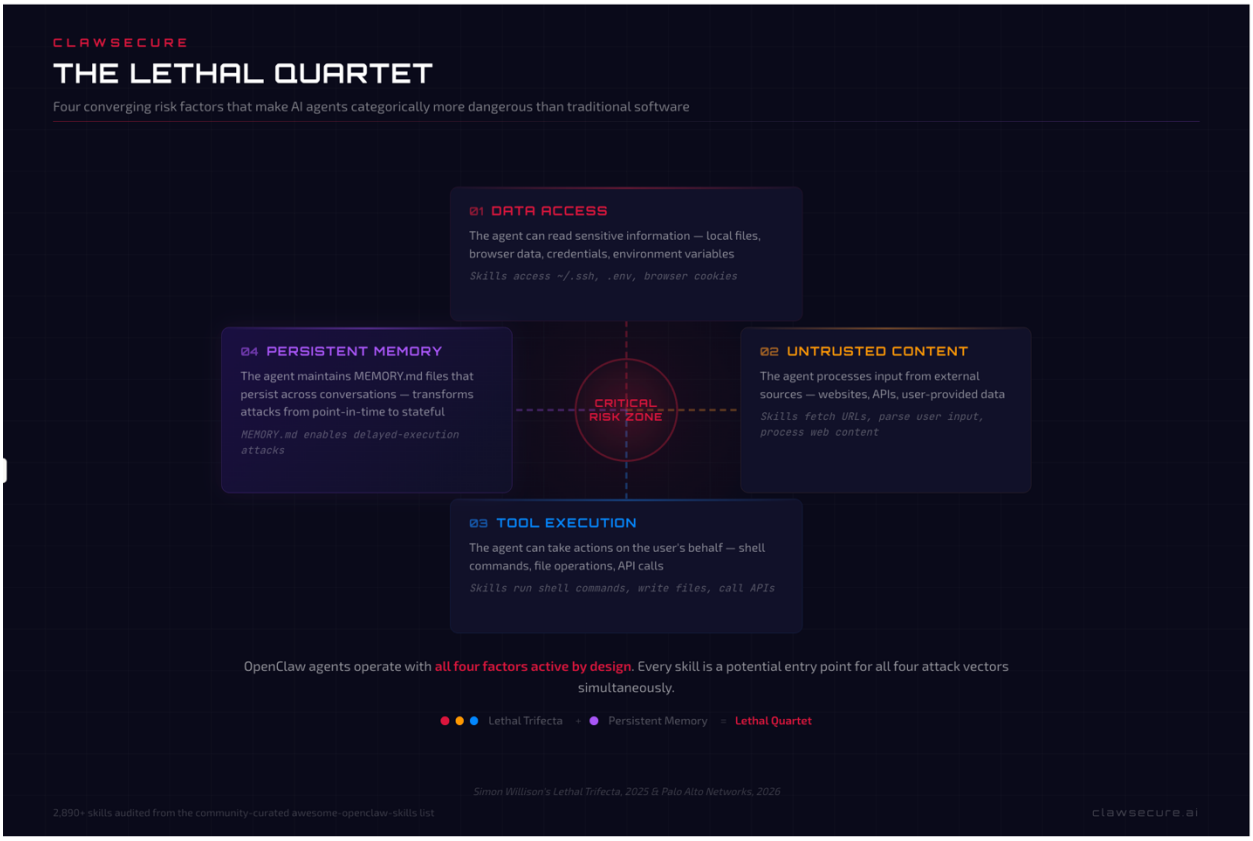
9. The Agentic Risk Model

The vulnerabilities documented in this report reflect fundamental architectural properties of agent-based systems that make AI agents categorically different from traditional software.

9.1 The Lethal Trifecta and the Lethal Quartet

Security researcher Simon Willison coined the term **"Lethal Trifecta"** to describe the convergence of three capabilities that makes AI agents uniquely dangerous: **Data Access, Untrusted Content,** and **Tool Execution.**

Palo Alto Networks expanded this into a **"Lethal Quartet"** by adding **Persistent Memory** as a fourth factor. OpenClaw agents, by design, operate with all four factors active.



9.2 Why This Changes Everything

A **command injection** in an agent with Data Access + Tool Execution + Persistent Memory is a potential pathway to persistent, undetectable compromise. A **prompt injection** in an agent with Tool Execution + Persistent Memory can result in actions taken on the user's behalf that persist

across sessions. **Credential harvesting** in an agent operating across multiple tools exposes the entire scope of the agent's access.

ClawSecure's 3-Layer Audit Protocol, Agentic Intent Verification, and Context-Aware Intelligence are designed specifically for this threat model, not the traditional software security model that generic scanners were built for.

10. Recommendations

10.1 For Users

Scan before you install. Every skill should be verified before installation. ClawSecure offers free, public scanning at clawsecure.ai. Look for ClawSecure Verified status (score 80+) as a baseline trust indicator.

Understand what you are granting access to. OpenClaw agents operate with significant system-level privileges. The near-total absence of permissions manifests (99.3% of skills lack config.json) means this assessment currently falls on the user.

Monitor for changes. ClawSecure's Watchtower system tracks skills continuously. Check a skill's current status before relying on an old scan result.

Be especially cautious with financial and crypto-themed skills. Our creator-level analysis identified patterns of consistently zero-scoring publishers targeting users interested in financial tools.

10.2 For Skill Creators

Include a config.json with every skill. Declare your skill's required permissions, capabilities, and resource needs. This is the single most impactful action creators can take to improve ecosystem trust.

Pin your dependencies. Unpinned version ranges create supply chain risk. Specify exact versions to prevent hijacking.

Never hardcode credentials. 41 skills in our dataset contain exposed API keys, tokens, or passwords.

Get scanned and earn Verified status. ClawSecure Verified status (score 80+) is a trust signal for your users.

Include a license. 82% of skills lack license declarations.

10.3 For Platforms and Marketplaces

Implement runtime integrity verification. ClawSecure's data shows skills change after publication: 35 modifications detected within 24 hours, and nearly 1 in 4 tracked skills (22.9%) have had their code modified since scanning. Platforms should integrate continuous verification.

Require permissions manifests. The absence of config.json across 99.3% of skills is an ecosystem-level failure that platforms are best positioned to mandate.

Implement publisher-level accountability. Patterns of consistently malicious publishing suggest that publisher verification and reputation systems are needed.

Integrate the Security Clearance API. ClawSecure's POST /api/v1/clearance endpoint enables platforms to verify agent integrity programmatically before granting access. The API returns real-time integrity status along with security scores and report links.

Adopt ecosystem-specific scanning. Generic scanners produce false positives when applied to agent skill code.

11. About ClawSecure

ClawSecure is the independent integrity layer for the agentic internet. We don't just scan files — we verify the soul of the agent as it evolves. We provide security scanning, runtime integrity verification, and continuous monitoring for AI agent skills, workflows, and multi-agent systems.

Our platform is built on five pillars of agentic security: Agentic Intent Verification, Agent-Native Auditing via our 3-Layer Audit Protocol, Watchtower for continuous Anti-Sleeper Protection, Context-Aware Intelligence, and the Identity Bridge for publisher accountability. Together, these capabilities cover 10 of 10 identified security categories for OpenClaw agents, the most comprehensive coverage available.

ClawSecure operates as a neutral, independent third-party auditor. We are not affiliated with, endorsed by, or in competition with any marketplace, platform, or agent ecosystem.

Core capabilities: Security Scanning (2,890+ skills audited, free at clawsecure.ai), Watchtower (continuous anti-sleeper integrity monitoring), Security Clearance API (programmatic real-time verification), Context-Aware Intelligence (ecosystem-specific threat analysis), Identity Bridge (tiered verification to KYC), Verified Agent Registry (curated directory of 80+ score skills).

About the Author

J.D. Salbego is the founder of ClawSecure. A 2x exited founder with over a decade of experience operating building across Web3, DeFi, and AI-driven autonomous financial systems, J.D. has scaled platforms to 310,000+ users and led operations processing over $4.5 billion in monthly trading volume. His career has increasingly focused on the intersection of autonomous systems, adversarial incentive design, and execution-layer security in capital-moving environments.

Contact: clawsecure.ai | @ClawSecure | @JDSalbego

Appendix: Definitions & Glossary

Agentic Intent Verification: ClawSecure's capability to verify whether an agent's actual code behavior matches its stated purpose.

Agent-Native Auditing: Security analysis purpose-built for agent architectures, as opposed to generic malware scanning.

C2 (Command and Control): Server infrastructure used by attackers to remotely control compromised systems.

ClawHavoc: A malware family targeting the OpenClaw ecosystem, originally discovered by Koi Security.

ClawSecure Verified: Status awarded to skills scoring 80 or above on ClawSecure's 3-Layer Audit Protocol.

Config.json: A configuration manifest file declaring a skill's required permissions. Currently absent from 99.3% of OpenClaw skills.

Context-Aware Intelligence: ClawSecure's ecosystem-specific threat analysis capability.

CVE (Common Vulnerabilities and Exposures): Standardized identifier for publicly known security vulnerabilities.

Hash Drift: A change in a skill's SHA-256 content hash between scans, indicating code modification.

Identity Bridge: ClawSecure's approach to linking code integrity with publisher accountability through tiered verification.

Lethal Trifecta: Framework coined by Simon Willison describing Data Access + Untrusted Content + Tool Execution. Expanded by Palo Alto Networks into the Lethal Quartet with Persistent Memory.

MEMORY.md: A file storing persistent agent memory across conversations.

Moltbook: A platform hosting over 2.2 million agent instances built on OpenClaw skills.

OpenClaw: An open-source platform for personal AI agents with over 180,000 users and 100,000 GitHub stars.

ReDoS: Regular Expression Denial of Service vulnerability.

Security Clearance API: ClawSecure's programmatic endpoint (POST /api/v1/clearance) for real-time integrity verification.

Sleeper Agent: A skill that passes initial scanning but is later modified to include malicious functionality.

SOUL.md: A file containing the OpenClaw agent's core personality, instructions, and operating parameters.

Vaccination Effect: The leverage dynamic where securing ~3,000 source skills protects 2.2M+ downstream agent instances.

Watchtower: ClawSecure's continuous anti-sleeper integrity monitoring system. Currently monitoring 2,880 skills 24/7.

This report was produced by ClawSecure (clawsecure.ai). Data reflects the state of the OpenClaw ecosystem as of February 2026. Scan results are not certifications. All third-party research is attributed to its original source. For questions, corrections, or media inquiries, contact us at clawsecure.ai or @ClawSecure on X/Twitter.