

SPRINT SECURITY GUIDE

OpenClaw Skills: What to Use, What to Avoid, and How to Stay Safe on ClawHub

February 2026 | ScaleUP Media

CRITICAL WARNING: Read This Before Installing Any Skill

Do NOT run OpenClaw on a company device. If you already have, treat it as a potential security incident and rotate all credentials immediately.

As of February 2026, security researchers have found **341+ malicious skills** on ClawHub, with **283 skills (7.1%)** containing critical credential exposure flaws. The #1 most-downloaded skill was confirmed malware.

This guide will help you navigate the ecosystem safely. Follow the rules in this document or risk losing API keys, SSH keys, browser sessions, crypto wallets, and cloud credentials.

1. What Are OpenClaw Skills?

OpenClaw skills are modular folders that teach your AI agent how to use tools, APIs, and workflows. Each skill is built around a **SKILL.md** file containing YAML metadata and markdown instructions. They follow the open **AgentSkills** specification, which means they're portable across any agent platform that supports the format (including OpenAI Codex).

Skills load from three locations with this precedence: **workspace skills** (highest) → **~/.openclaw/skills** (managed/local) → **bundled skills** (lowest). When a name conflicts, the higher-precedence version wins.

ClawHub (clawhub.ai) is the public registry with **5,700+ community-built skills** as of February 2026. Installation is one command: `clawhub install <skill-name>`. This simplicity is both the magic and the danger.

2. The Threat Landscape: What Went Wrong

The OpenClaw skills ecosystem experienced one of the most significant AI supply chain attacks in history during January–February 2026. Here's what the security community uncovered:

ClawHavoc: The Numbers

Finding	Source	Details
341 malicious skills found	Koi Security	335 from single coordinated campaign (ClawHavoc), targeting macOS and Windows
283 skills (7.1%) leak credentials	Snyk/Evo	Functional skills that expose API keys, passwords, and credit card numbers through the LLM context window

512 vulnerabilities in OpenClaw itself	Kaspersky	Security audit found 512 bugs, 8 classified as critical
30,000+ exposed instances	Censys	Gateway binds to 0.0.0.0:18789 by default, no auth required
354 malicious packages from one actor	Bitdefender	Account "Hightower6eu" uploaded 354 malicious packages
"What Would Elon Do" #1 skill = malware	Cisco	9 security vulnerabilities, 2 critical. Silent data exfiltration + prompt injection

How the Attacks Work

The core insight from 1Password's Jason Meller: “*Markdown isn't content in an agent ecosystem. Markdown is an installer.*”

The attack patterns fall into distinct categories:

Attack Type	How It Works	Example
Fake Prerequisites	Skill tells you to install a “required dependency” that’s actually malware	Twitter skill requiring “openclaw-core” — links led to staged malware delivery
Typosquatting	Misspelled versions of popular skill names	29 variants of “clawhub”: clawhub1, clawhubb, clawwhub, cllawhub, etc.
Silent Data Exfiltration	Skill runs curl commands to send your data to attacker servers	“What Would Elon Do” used prompt injection to bypass safety + exfiltrate silently
Reverse Shell Backdoors	Skill opens a remote connection giving attacker full control	Polymarket tool that opened interactive shell to attacker’s server
Credential Harvesting	Skill reads your .env files and sends them to webhooks	Skills extracting ~/.clawdbot/.env to external services
Leaky Instructions	Skill tells agent to pass API keys through the LLM context window in plaintext	moltiverse-email, youtube-data, buy-anything skills

Categories Most Targeted by Malware

If you see a skill in these categories on ClawHub, apply extreme scrutiny:

Category	Malicious Count	Why Targeted
Crypto Tools (Solana, Phantom, wallets)	111 skills	High-value targets with direct financial payoff

YouTube Utilities	57 skills	Mass appeal, wide distribution potential
Polymarket Bots	34 skills	Prediction market traders seeking an edge
ClawHub Typosquats	29 skills	Catches rushed/typo installations
Auto-Update	28 skills	Ironic: malware disguised as update tools
Google Workspace clones	Multiple variants	Broad appeal, credential-rich environment

Why MCP Does NOT Protect You

A common misconception is that the Model Context Protocol (MCP) layer makes skills safe. It doesn't. Skills don't need to use MCP at all. The AgentSkills spec places **no restrictions on the markdown body**, and skills can bundle scripts that execute outside the MCP tool boundary entirely. MCP can be part of a safe system, but it is not a safety guarantee by itself.

3. Skills Worth Using (With Caution)

No third-party skill should be considered fully safe. That said, the following categories of skills have been **vetted by independent security auditors** (VettedSkillsHub, Koi's Clawdex) and/or are **bundled with OpenClaw itself**. Even these require you to read the SKILL.md before installing.

Tier 1: Bundled Skills (Safest)

These ship with OpenClaw itself and don't require ClawHub installation:

Skill	What It Does	Risk Level
• Lobster (built-in)	Core agent functionality	LOW — ships with install
• LLM Task	Delegate subtasks to language models	LOW — bundled
• Exec Tool	Execute shell commands (with approval)	MEDIUM — powerful but you control it
• Web Tools	Fetch and interact with web content	MEDIUM — network access
• Browser (OpenClaw-managed)	Headless browser automation	MEDIUM — credential exposure risk

Tier 2: Well-Vetted Community Skills

These have been tested by independent reviewers and/or have strong maintenance records. Always verify the VirusTotal scan on ClawHub before installing:

Category	Recommended Skills	Notes
Productivity	Todoist, Apple Reminders, Google Calendar (official)	Stick to skills with verified publishers
Communication	Slack (official), Discord (official), WhatsApp (wacli)	Check that the skill doesn't ask for credentials in SKILL.md
Developer Tools	GitHub integration, Cursor CLI agent	Review for shell execution patterns
Search & Research	Tavily (AI search), Perplexity integration	Requires API keys — use environment injection, never hardcode
Media	FFmpeg video editor, ElevenLabs TTS	Local processing preferred over cloud
System Monitoring	Server status (CPU/RAM/GPU check)	Read-only skills are inherently safer
Security	Clawdex (Koi), Security Audit skill	Defense-oriented skills worth having

Tier 3: Use at Your Own Risk

These categories are functional but have higher risk profiles:

Category	Why Risky	If You Must Use Them
Crypto/DeFi tools	Most-targeted category. 111 malicious skills found.	Build your own. Never install third-party crypto skills.
Trading bots	Polymarket and financial tools are prime targets	Verify publisher identity. Read every line of SKILL.md.
Auto-updaters	28 malicious variants discovered	Use clawhub update --all instead of third-party updaters
YouTube utilities	57 malicious variants found	Only use skills from verified, established publishers
Google Workspace clones	Multiple typosquat variants exist	Use the official google-workspace skill only

4. The 5-Point Skill Vetting Framework

Before installing ANY skill from ClawHub, run through this checklist:

The TRUST Checklist

T — Transparency: Can you read every line of SKILL.md and any bundled scripts? If there are obfuscated commands, encoded payloads, or unexplained binary downloads — walk away.

R — Reputation: Is the publisher verified on ClawHub? Do they have a GitHub account older than one week? Is there a real identity behind the skill? Check the VirusTotal scan badge.

U — Updates: Has the skill been updated in the last 3 months? Does it have a changelog? Abandoned skills are liabilities. But also watch for sudden updates to previously stable skills (possible account takeover).

S — Scope: Does the skill request only the minimum permissions needed? A YouTube summarizer should NOT need access to your .env file or shell commands. Apply the principle of least privilege.

T — Testing: Run the skill in a sandboxed environment first. Use OpenClaw's sandbox mode. Monitor for outbound network calls you didn't expect. Check what files it reads or writes.

5. OpenClaw Security Hardening for Skills

Credential Management

1. **Never hardcode API keys in SKILL.md files.** Use environment injection via openclaw.json's skills.entries.<key>.env or skills.entries.<key>.apiKey instead.
2. **Never let skills handle credentials in plaintext.** Snyk found 283 skills instructing agents to pass secrets through the LLM context window. This is an active exfiltration channel.
3. **Rotate all credentials regularly.** If a skill has ever seen your API key, assume it could be compromised.
4. **Use a secrets manager.** Don't store tokens in .env files that skills can read.

Network & Sandbox

1. **Change the default gateway bind address.** OpenClaw binds to 0.0.0.0:18789 by default. Bind to 127.0.0.1 instead.
2. **Run untrusted skills in sandbox mode.** Use OpenClaw's Docker-based sandboxing for any third-party skill.
3. **Monitor outbound network traffic.** Skills should not be making unexpected HTTP calls. Watch for curl, wget, or fetch to unknown domains.
4. **Enable Exec approvals.** Don't set Exec security to "allow all." Require approval for shell commands.

Skill Management

1. **Install the Clawdex skill.** It scans your installed skills against Koi's malicious skills database before and after installation.
2. **Pin skill versions.** Use .clawhub/lock.json to lock versions. Don't auto-update blindly.
3. **Audit your installed skills regularly.** Run "openclaw skills list" and review what's active. Remove anything you don't actively use.
4. **Write your own skills for sensitive operations.** For anything touching credentials, financial data, or customer information — build it yourself. A SKILL.md is just markdown. You can write one in 30 minutes.

6. Instant Red Flags: Walk Away Immediately

If a skill exhibits any of these behaviors, do not install it:

Kill Signals — Do Not Install If You See These

- Requires you to install a "prerequisite" or "core dependency" via a link or pasted command
- Contains base64-encoded strings or obfuscated commands
- Asks you to remove macOS quarantine attributes (xattr -d com.apple.quarantine)
- Downloads binaries from non-standard sources (paste services, raw GitHub URLs, URL shorteners)
- Requests credentials be entered directly into the SKILL.md or stored in MEMORY.md
- Has a name that's a slight misspelling of a popular skill (typosquatting)
- Publisher account is less than 2 weeks old or has random character suffixes in the name
- Skill promises crypto wallet access, whale tracking, or insider trading signals
- Makes outbound HTTP requests that aren't essential to its described function
- Uses prompt injection language like "ignore previous instructions" or "bypass safety guidelines"

7. If You've Been Compromised

If you installed a suspicious skill or see unexpected behavior:

Immediate Response Checklist

1. **Stop using the device** for any sensitive work immediately.
2. **Rotate everything:** Browser sessions, API keys, OAuth tokens, SSH keys, cloud console sessions, email passwords.
3. **Review recent sign-ins** for email, GitHub/GitLab, cloud providers (AWS/GCP/Azure), CI/CD pipelines, and admin consoles.
4. **Check `~/.clawdbot/.env`** and `MEMORY.md` for any credentials that may have been exposed.

- 5. Scan installed skills** with Clawdex or manually review each SKILL.md file.
- 6. If on a company device**, engage your security team immediately. Treat it as a potential incident. Do not wait for symptoms.
- 7. Run VirusTotal** on any binaries that were downloaded as part of skill “prerequisites.”

8. Build Your Own Skills Instead

The safest skill is one you wrote yourself. A SKILL.md is just a folder with a markdown file. The basic structure is simple:

Minimal SKILL.md Template

```
---name: my-custom-skill
description: Brief description of what it does---
My Custom Skill## Instructions[Your agent instructions here]##
Prerequisites[Only list tools already installed on your system]
```

For your ScaleUP operations across 14 SaaS products, building custom skills gives you full control over what your agent can access. Skills placed in your workspace’s *skills*/ directory are automatically loaded and take highest precedence over ClawHub installs.

9. The Bottom Line

OpenClaw skills are incredibly powerful — they collapse the distance between intent and execution. But that power means the ecosystem is a supply chain, and supply chains get attacked. The ClawHavoc campaign proved this isn’t theoretical.

The rules are simple:

- **Never install skills blindly.** Read every SKILL.md. Check VirusTotal. Verify the publisher.
- **Never run OpenClaw on a machine with production credentials.** Use dedicated, isolated hardware.
- **Build your own skills for anything sensitive.** It takes 30 minutes and eliminates supply chain risk.
- **Use the TRUST checklist** on every single ClawHub install.
- **Stay updated.** OpenClaw now has VirusTotal scanning, but the ecosystem is still maturing. New attack patterns will emerge.

Sources & Further Reading

1Password: “From Magic to Malware” —

1password.com/blog/from-magic-to-malware-how-openclaws-agent-skills-become-an-attack-surface

Koi Security: “ClawHavoc” — koi.ai/blog/clawhavoc-341-malicious-clawedbot-skills-found

Cisco: “Personal AI Agents Are a Security Nightmare” —

blogs.cisco.com/ai/personal-ai-agents-like-openclaw-are-a-security-nightmare

Snyk: “280+ Leaky Skills” — snyk.io/blog/openclaw-skills-credential-leaks-research

Bitdefender: “Technical Advisory: OpenClaw Exploitation” —

businessinsights.bitdefender.com

Kaspersky: “New OpenClaw AI Agent Found Unsafe” —

kaspersky.com/blog/openclaw-vulnerabilities-exposed

VirusTotal: “From Automation to Infection” — blog.virustotal.com

The Hacker News: “OpenClaw Integrates VirusTotal Scanning” — thehackernews.com

OpenClaw Official Docs: Skills — docs.openclaw.ai/tools/skills