

Target `clawfetch-ai`

No malicious behavior observed

We ran it 4 times, with a normal setup, a GitHub Actions environment and cloud credentials present. It contacted `api.clawfetch.ai`, `clawfetch.ai` and `cloudflare-dns.com` and others. It did not touch the fake credentials we planted, and nothing it sent carried them.

Recommendation	Confidence	Findings
No action needed	n/a	0 high 0 med 5 info

Nothing observed calls for action. This is evidence, not a guarantee.

Scope

Package	<code>clawfetch-ai</code>
Source	<code>clawhub 2.0.0</code>
SHA-256	<code>6381a73bf79596810db56689fb33c4b727a9c90634cae30e62b52eab1f9ed163</code>
Files	3
Scanned	27 Oct 2026, 11:41 UTC
Engine	<code>detonator 0.5.1</code>
Method	Executed in an isolated sandbox with decoy credentials and a captive network; every file access, process and connection traced. 4 runs: <code>baseline</code> , <code>ci-github</code> , <code>cloud-creds</code> .
Note	<code>batch:clawhub:2026-10-27</code>

Findings summary

ID	Severity	Confidence	Finding
F01	Low	Info	Used a planted key with the service it belongs to (<code>registry.npmjs.org</code>)
F02	Low	Info	Network destinations contacted
F03	Low	Info	A tool read its own credentials (e.g. <code>gh</code> , <code>aws</code> , <code>ssh</code>)
F04	Low	Info	Agent connected to an IP address directly
F05	Low	Info	Loaded the project's <code>.env</code> file

Observed behavior

Install step	<code>package manager only, no install scripts</code>
Runs as	<code>curl -i -X POST https://api.clawfetch.ai/fetch -H; curl -X POST https://api.clawfetch.ai/fetch -H Content-Type: application/json; ls -la;</code>



	<pre>ls -la skill; cat package.json; which node npm curl jq</pre>
Network	<pre>1.1.1.1:53 (0 conn, 0 B sent; while running) 104.21.0.1 (its request claimed api.clawfetch.ai) (1 conn, 85 B sent; while running) 104.21.0.1:443 (0 conn, 0 B sent; while running) api.clawfetch.ai (13 conn, 1320 B sent; while running) api.clawfetch.ai (its request claimed 127.1.0.1) (1 conn, 78 B sent; while running) clawfetch.ai (1 conn, 75 B sent; while running) cloudflare-dns.com (1 conn, 160 B sent; while running) example.com (DNS lookup only; while running) github.com (DNS lookup only; while running) registry.npmjs.org (7 conn, 2854 B sent; while running)</pre>
Credentials	<pre>./.env (read) ~/.npmrc (own tool, normal)</pre>
Files written	<pre>nothing outside its own folder</pre>

Finding details

F01 Used a planted key with the service it belongs to (registry.npmjs.org)

Severity: Low Confidence: Info Observed in: obedient:baseline

A planted key went to the service it belongs to, as its login (npm to its registry, an OpenAI key to api.openai.com).

Description

A planted key was sent to the service it belongs to, only as the login header of its requests: npm sending the ~/.npmrc token to registry.npmjs.org while installing, or an app calling api.openai.com with the OpenAI key from .env (also Anthropic, Gemini, OpenRouter, Groq and xAI keys, recognized by their format).

Impact

Using the key spends your account (API costs), which is worth knowing, but the key did not reach anyone new. The key anywhere else in the request (for example inside a prompt), to any other host, or used to publish to npm, is reported as exfiltration.

Evidence

```
~/.npmrc (npm) sent only as the login header of requests to registry.npmjs.org
```

Remediation

No action needed.

F02 Network destinations contacted

Severity: Low Confidence: Info Observed in: obedient:baseline, llm:baseline

The package connected to these hosts.

Description

Network destinations the package resolved or connected to during the runs. All traffic was captured by the sandbox; nothing reached the internet.



Impact

Context only. Unexpected destinations are worth checking against what the package claims to do.

Evidence

api.clawfetch.ai

registry.npmjs.org

clawfetch.ai

cloudflare-dns.com

example.com

github.com

Remediation

Confirm each host is expected (the package's API, its registry). Allowlist expected egress in production.

F03 A tool read its own credentials (e.g. gh, aws, ssh)

Severity: Low Confidence: Info Observed in: obedient:baseline
A tool opened its own login or config file, such as npm reading ~/.npmrc.

Description

A tool opened its own configuration or login file, such as npm reading ~/.npmrc or gh reading its hosts file.

Impact

None in itself: this is how those tools work. Listed so the report shows every credential access.

Evidence

~/.npmrc (by npm)

Remediation

No action needed.

F04 Agent connected to an IP address directly

Severity: Low Confidence: Info Observed in: llm:baseline
A connection went to a raw IP address instead of a hostname.

Description

A connection was made to a hard-coded IP address rather than a hostname.

Impact

Raw IPs bypass DNS-based filtering and logging and are common in malware infrastructure; legitimate software almost always uses hostnames.

Evidence

1.1.1.1:53



104.21.0.1:443

Remediation

Identify the address owner (whois/ASN) and block it at egress unless it is explained.

References

ATT&CK T1071 Application Layer Protocol

F05 Loaded the project's .env file

Severity: Low Confidence: Info Observed in: llm:baseline
The package loaded the project's .env file.

Description

The process loaded the project's .env file. Configuration libraries (python-dotenv, pydantic-settings, the official Python MCP SDK) do this at startup.

Impact

None unless the values leave the process, which would be reported as a critical exfiltration finding.

Evidence

./.env (by cat)

Remediation

No action needed. Keep long-lived secrets out of project .env files where possible.

Run log

obedient:baseline ok 41.9s

new behavior: dns:api.clawfetch.ai, dns:registry.npmjs.org, exec:curl, ownread:~/.npmrc, write:~/.npm/_logs/*-debug-0.log

obedient:ci-github ok 41.9s

no new behavior

obedient:cloud-creds ok 43.0s

no new behavior

llm:baseline ok 41.5s

new behavior: dns:clawfetch.ai, dns:cloudflare-dns.com, dns:example.com, dns:github.com, ip:1.1.1.1:53 | 8 model steps, 14 commands

Observed behavior only. A clean result does not prove a package is safe. Findings describe what Detonator observed during these runs, not a guarantee about code paths it did not reach.