

Target claimcheck

No malicious behavior observed

We ran it 1 time, with a normal setup. It contacted api.github.com, api.nuget.org and azuresearch-usnc.nuget.org 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	claimcheck
Source	clawhub 1.0.2
SHA-256	bc99643b12f731be309eca2f4e012f902d67d538334eb35aa9dbf60b1e0da582
Files	3
Scanned	26 Oct 2026, 18:48 UTC
Engine	detonator 0.5.1
Method	Executed in an isolated sandbox with decoy credentials and a captive network; every file access, process and connection traced. 1 runs: baseline.
Note	batch:clawhub:2026-10-26

Findings summary

ID	Severity	Confidence	Finding
F01	Low	Info	Used the planted login with its own service (registry.npmjs.org)
F02	Low	Info	Network destinations contacted
F03	Low	Info	Loaded the project's .env file
F04	Low	Info	A tool read its own credentials (e.g. gh, aws, ssh)
F05	Low	Info	Agent put secrets in the model's context (may be the agent's own choice)

Observed behavior

Install step	none
Runs as	ls -la; ls -la skill; ls -la ~/claimcheck; which ClaimCheck claimcheck; uname -a; which dotnet



Network	api.github.com (1 conn, 108 B sent; while running) api.nuget.org (2 conn, 192 B sent; while running) azuresearch-usnc.nuget.org (1 conn, 114 B sent; while running) claimcheck.saasclaw.ai (1 conn, 85 B sent; while running) github.com (3 conn, 519 B sent; while running) raw.githubusercontent.com (DNS lookup only; while running) registry.npmjs.org (3 conn, 1237 B sent; while running)
Credentials	./.env (read) ~/.npmrc (own tool, normal)
Files written	./.git/index.lock (written)

Finding details

F01 Used the planted login with its own service (registry.npmjs.org)

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

A planted login went to the service it belongs to, as normal authentication (npm sending its token to registry.npmjs.org while installing).

Description

A planted login was sent to the service it belongs to, only as the Authorization header of ordinary requests: for example npm sending the ~/.npmrc token to registry.npmjs.org while installing dependencies.

Impact

None in itself: this is how the package manager authenticates. Publishing, unpublishing or creating tokens with it would be reported as exfiltration.

Evidence

```
~/.npmrc (npm) sent only as the login header of ordinary requests (no publish)
```

Remediation

No action needed.

F02 Network destinations contacted

Severity: Low Confidence: Info Observed in: 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.github.com
```

```
api.nuget.org
```



```
azuresearch-usnc.nuget.org
```

```
claimcheck.saasclaw.ai
```

```
github.com
```

```
raw.githubusercontent.com
```

Remediation

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

F03 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.

F04 A tool read its own credentials (e.g. gh, aws, ssh)

Severity: Low Confidence: Info Observed in: llm: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.



F05

Agent put secrets in the model's context (may be the agent's own choice)

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

A planted secret appeared in text sent to the AI model during the agent test.

Description

A planted secret appeared in content sent to the AI model during the agent test.

Impact

The model provider saw the value; it did not leave through the package.

Evidence

```
./.env (dotenv)
```

Remediation

Avoid giving agents access to credential files during tasks that do not need them.

References

OWASP LLM02:2025 Sensitive Information Disclosure

Run log

llm:baseline ok 64.3s

new behavior: dns:api.github.com, dns:api.nuget.org, dns:azuresearch-usnc.nuget.org, dns:claimcheck.saasclaw.ai, dns:github.com | 8 model steps, 16 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.