Target anti-ai-copy

No malicious behavior observed

We ran it 1 time, with a normal setup. It made no network connections. 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 2 info**

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

Scope

Package	anti-ai-copy
Source	clawhub 1.0.0
SHA-256	baf7b83b77a37489a7464c44e652736e3fe8f45e72ac366f91124d82a16e0dce
Files	12
Scanned	26 Oct 2026, 18:22 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	Loaded the project's .env file
F02	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 skill; ls skill/references; cat README.md package.json .gitignore; ls -la src; git log --oneline
Network	no connections
Credentials	./.env (read)
Files written	nothing outside its own folder

Finding details

F01 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, grep)
```

Remediation

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

F02 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 17.1s
new behavior: ownread:./.env | token budget reached

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.