Target reply-to-every-review

No malicious behavior observed

We ran it 1 time, with a normal setup. It contacted api.appgoat.ai. 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 3 info**

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

Scope

Package	reply-to-every-review
Source	clawhub 1.0.0
SHA-256	3fc93e3217e8f65f04ad6b7ea2075256bf6779df5efeb9a5a5b1101f8e58049a
Files	4
Scanned	27 Oct 2026, 12:25 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:2026-10-27

Findings summary

ID	Severity	Confidence	Finding
F01	Low	Info	Network destinations contacted
F02	Low	Info	Loaded the project's .env file
F03	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; head -100; cat skill/reference/api.md; env; grep -i -E appgoat|anthropic|api_key|mcp
Network	api.appgoat.ai (3 conn, 307 B sent; while running)
Credentials	./.env (read)

Files written	`nothing outside its own folder`

Finding details

F01 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.appgoat.ai
```

Remediation

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

F02 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.

F03 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 16.4s

new behavior: dns:api.appgoat.ai, exec:curl, ownread:./.env | 6 model steps, 6 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.