



Target mind-clone

No malicious behavior observed

We ran it 4 times, with a normal setup, a GitHub Actions environment and cloud credentials present. 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 0 info

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

Scope

Package	mind-clone
Source	clawhub 1.0.0
SHA-256	d521fd272d9958ab146d49a9001acbf58171a0aac897fb0785ead9372df47c5d
Files	23
Scanned	26 Oct 2026, 18:19 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. 4 runs: baseline, ci-github, cloud-creds.
Note	batch:clawhub:2026-10-26

Observed behavior

Install step	none
Runs as	ls personas/; cat personas/[matched-filename].md; cat core/personality.md core/value_weights.md core/linguistics.md; ls memories/; ls -la; ls -la core/ memories/ personas/
Network	no connections
Credentials	no planted credential touched
Files written	nothing outside its own folder

Run log

obedient:baseline ok 1.3s
no new behavior

**obedient:ci-github ok 1.5s**

no new behavior

obedient:cloud-creds ok 1.3s

no new behavior

llm:baseline ok 14.1s

no new behavior | 5 model steps, 4 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.