

Target refining-the-tone-of-ai

No malicious behavior observed

We ran it 8 times, with a normal setup, LC_MESSAGES set, docs:draft.md --profile general and a GitHub Actions environment. 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	refining-the-tone-of-ai
Source	clawhub 0.1.0
SHA-256	d32cfe297f173ae1e3e30441f6f3b836cab5ba377b3b9e4181c181ede9d5f02
Files	30
Scanned	26 Oct 2026, 18:41 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. 8 runs: baseline, guided:LC_MESSAGES, docs:draft.md --profile general, ci-github.
Note	batch:clawhub:2026-10-26

Observed behavior

Install step	none
Runs as	python3 <skill>/scripts/audit_ai_flavor.py; python3 <skill>/scripts/audit_ai_flavor.py draft.md --profile general; python3 <skill>/tests/test_audit_ai_flavor.py; /usr/bin/python3 <skill>/scripts/audit_ai_flavor.py --stdin --json --profile general; /usr/bin/python3 <skill>/scripts/audit_ai_flavor.py <skill>/tests/fixtures/*_after.md; /usr/bin/python3 <skill>/scripts/audit_ai_flavor.py --stdin --fail-on-review
Network	no connections
Credentials	no planted credential touched
Files written	nothing outside its own folder



Run log

scripts:baseline ok 2.5s

no new behavior

scripts:guided:LC_MESSAGES ok 1.3s

no new behavior

scripts:docs:draft.md --profile general ok 1.9s

no new behavior

scripts:baseline ok 18.4s

new behavior: write:./draft.md, write:./nested, write:./note.md, write:./skip.py, write:./terms.txt

scripts:guided:LC_MESSAGES ok 15.0s

no new behavior

scripts:ci-github ok 14.5s

no new behavior

scripts:baseline ok 2.7s

no new behavior

scripts:guided:LC_MESSAGES ok 2.5s

no new behavior

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.