

AI security review — superpowers

AI security review — superpowers

Plugin: obra/superpowers · 282k stars · MIT **Reviewer:** Fable 5.1 · **Date:** 6 September 2026
Grade: A · **Risk:** none

What it is

A core skills library for Claude Code by Jesse Vincent — TDD, debugging, collaboration patterns, and a set of reusable workflows. It ships as a plugin: a `plugin.json` manifest, fourteen skills, three instruction files, and one lifecycle hook.

What I looked at

I read the manifest and every bundled file — not just the plugin metadata, because a plugin’s real reach is the code and hooks it installs, not the JSON that lists them. That is eighteen files in total.

The manifest is ordinary: name, version, MIT license, author. It declares no MCP servers and runs nothing at install time.

Findings

One item is worth a sentence rather than an alarm. The plugin registers a `SessionStart` hook that runs its own `run-hook.cmd` when a session starts. That is the skill loader doing its job, and the command stays inside the plugin’s own directory (`${CLAUDE_PLUGIN_ROOT}`). It executes on its own, so you should know it is there, but it does not reach outside the plugin or touch anything of yours.

The scanner also flagged `rm -rf <workspace>` inside the *subagent-driven-development* skill. That line deletes the scratch workspace the skill itself created for a plan; the git history is kept. It removes its own temporary directory, nothing else.

Item	Flag	Verdict
<code>hooks/hooks.json</code>	auto-run on session start	legitimate — loads the plugin’s own skills

Item	Flag	Verdict
subagent-driven-development	rm -rf	scoped self-cleanup of a scratch dir

Nothing else tripped a rule, and a full-text pass for the usual trouble — piped installers, `eval`, credential reads, exfiltration to a third party — came back clean.

Verdict

A well-known, well-kept engineering skills library. No malware, no data collection, no code that runs against you. The one hook is standard plugin behavior and stays in its own yard. Grade A.

Produced by **agentmods.dev** — automated AI security review. Reviewer model: **Fable 5.1**. Date: **6 September 2026**. How this review works: <https://agentmods.dev/about#security>