NVIDIA SkillSpector: Check if an AI Agent Skill Is Safe Before You Install It

Kabui, Charles

2026-06-29

Read at ToKnow.ai

NVIDIA released SkillSpector, a free, open-source scanner that checks whether an AI agent skill is safe to install. A skill is a bundle of plain instructions plus real executable code that extends an agent like Claude Code, Codex CLI, or Gemini CLI, and it runs with your access and credentials. It scans a folder, file, zip, or Git repo for 68 vulnerability patterns across 17 categories, among them prompt injection, data exfiltration, and privilege escalation. It runs a

fast automated read of the code, then an optional pass where a language model judges intent and looks up known security flaws (CVEs) through OSV.dev, and returns a risk score from 0 to 100.

Skills are spreading faster than anyone can vet them. A study that analyzed 31,132 agent skills from two marketplaces found 26.1% contained at least one vulnerability and 5.2% showed likely malicious intent, with script-bundling skills 2.12 times more likely to be vulnerable than instruction-only ones. SkillSpector turns a manual security review into one command that runs in seconds, and it can sit inside your agent as a gate that blocks a risky skill before it installs.

This changes how the skills supply chain gets secured. For a year the advice was to install only skills you trust, which fails when popular packages can be the infected ones. A scanner from a major vendor moves vetting from a judgment call to an automated check, the direction OWASP pushes in its Agentic Skills Top 10. It answers the supply-chain risk ToKnow.ai documented at ClawHub.

Sources:

- NVIDIA SkillSpector on GitHub
- Agent Skills in the Wild: An Empirical Study of Security Vulnerabilities at Scale (arXiv)
- OSV.dev open-source vulnerability database
- OWASP Agentic Skills Top 10
