



Agentic AI Foundation

AGNTCon + MCPCon North America

Where the **agentic stack** is being built.

Shipping Agent Skills Safely: CI/CD, Evals, and Guardrails

Michael Larson, Christensen Solutions

Who Am I?

- Michael Larson
 - 19+ years working on large scale systems
 - Developer platforms
 - High-reliability software
 - AI assisted engineering
 - MCP servers
 - Agent-based tooling



Skills are everywhere

3.9M

find-skills installs

992K

frontend-design installs

Who has already installed skills?

skills.sh displayed totals, rounded. Snapshot: October 15, 2026

Problem 1: “It worked once”



Oren Melamed 

@OrenMe

...

Watched a team ship a “great” skill last month.
It worked in the demo.
It quietly got worse every time the model changed underneath it.
No evals. No version. No way to say “v12 is worse than v11, roll it back.”
We would never treat a library like that.
We treat prompts and skills like group-chat folklore.
That’s something which I believe will become a big problem.

5:56 AM · Sep 21, 2026 · **999** Views

Problem 2: Malicious attacks

Skills are already being used in attacks

Attackers disguise malware as setup steps

Encrypted archives evade content scans

Problem 3: Cost

 Forbes

Subscribe: Less than \$2/wk

5+

EDITORS' PICK | INNOVATION > CLOUD

Uber Burns Its 2026 AI Budget In Four Months On Claude Code

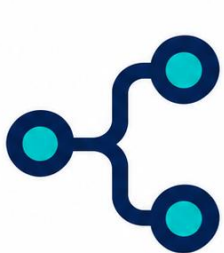
By [Janakiram MSV](#), Senior Contributor. ⓘ I cover emerging technologies wit... 

[Follow Author](#)

Published May 17, 2026, 11:18pm EDT, Updated May 27, 2026, 12:17pm EDT

The solution

Treat skills like software



Version



Evaluate



Review



Release

Setting up evals

Recommended recipe:

1. Collect 5 to 10 tasks plus known failure modes
2. Write deterministic assertions
3. Compare current and proposed skills in CI
4. Re-run the suite when the model changes
5. Add adversarial cases

A paired eval in practice



Same task

A fixture with an off-by-one error and a missing null check



Compare versions

Current skill (v1) and proposed skill (v2)



Check behavior

Find known defects, avoid extra findings, preserve files



Run in CI

Repeat on skill and model changes

My TDD harness setup

Project setup

1. Install the harness
2. Run tdd-setup
Learn the project's test conventions
3. Select the tdd agent

Skills handle planning, phase rules and test runs

Promptfoo scores runs in disposable fixture repos

Two agents

tdd

Plans one behavior and delegates a cycle

tdd-cycle-runner

Runs one loop in its own context
Returns a JSON contract

A Promptfoo test

promptfooconfig.yaml (excerpt)

tests:

- threshold: 1

vars:

fixtureId: calculator-cli-go

taskPrompt: >-

Add a subtract subcommand
to the calc CLI using TDD.

assert:

- type: javascript

value: file://./assertions/commit-discipline.js

Fails when code changes remain uncommitted

Jev in my evals

Real results from my TDD harness



Before / after code
Test execution output
Git diffs and commits

state: evidence,
questions: questionsForPhase(phase)

Recorded RED call: 222 ms

RED judgment	Jev result
Focused test	1.95 / 2
Intended failure	0.96 P(yes)
Scope preserved	0.96 P(yes)
No RED commit	0.91 P(yes)

jev-1.14.0, one recorded call

Total Jev spend so far: \$0.01 across all my runs

Promptfoo records scores. CI checks completion. No pass thresholds.

Copilot TDD harness

Evaluation pipeline

Jev live evaluation

Latest evaluation results

Published scorecard and Promptfoo report

github.com/mrlarson2007/copilot-tdd-harness

Distributing skills like packages



APM



skills.sh + skills CLI

Portability needs testing

Example: a skill assumes a tool named Bash

Another agent may expose a different shell interface
or require different approvals

Run the same eval suite on each supported agent

Checking skills for attacks

Scan



Review



Sandbox



**Add an eval to one skill
you're sharing**

Questions?



Agentic AI Foundation

AGNTCon + MCPCon

North America

Where the **agentic stack** is **being built**.

Thank you

Contact me

Michael Larson

Christensen Solutions

mlarson@christensensolutions.com



Scan to connect on LinkedIn