The agent said done.
What should you verify?

A practical guide for engineering and AI-platform leaders deciding whether recurring, unattended, or high-fan-out coding-agent workflows are ready to scale.

0 / 20

successful write results in one recurring job

Every run had a transcript. None recorded a successful write result. This one-machine observation is the prompt for the guide, not an industry benchmark.

SCOPE / 1 MACHINE / 60 DAYS / 1,166 RUNS / NOT A BENCHMARK

Use this guide to answer three questions

1. **Which agent workflow matters enough to examine?**
2. **What delivery evidence exists - and where does it stop?**
3. **What is the smallest operating decision the evidence can support?**

02

BEFORE SCALING

Five questions before you scale

Use one important workflow as the unit of analysis

01

What is success?

Name the expected file, test, commit, pull request, deployment, or reviewed artifact before examining the run.

02

Who checks it?

Decide whether the agent, CI, a delivery probe, or a person verifies the expected evidence.

03

What is missing?

Distinguish a failed run, absent evidence, an unavailable remote target, and an unresolved outcome.

04

What did it cost?

Include child-agent volume and cache-read tokens rather than looking only at generated output.

05

What can change?

Tie the finding to a workflow, agent definition, permission, model, context policy, or probe.

GOOD FIRST CANDIDATE

A recurring, unattended, repository-wide, or high-fan-out workflow whose output affects a real engineering decision.

03

EVIDENCE

Use an evidence ladder

Each step answers a stronger question than the one before it

1	**Transcript ended** Did the recorded interaction reach a terminal state?	SUPPORTED
2	**Tool reported success** Did Write, Edit, or NotebookEdit record a successful result?	SUPPORTED
3	**Local target present now** Can the recognized successful target be found on this filesystem now?	SUPPORTED
4	**Commit or PR attributed** Can repository evidence safely credit this run?	NEXT PROBE
5	**Deployment or external action** Did work reach the external system it promised?	NOT BUILT
6	**Semantic quality** Is the artifact correct, useful, and accepted?	HUMAN / TEST

RULE

Never promote an earlier rung into a later claim. A clean transcript is not a commit; a present file is not proof that it is unchanged or correct.

04

PRIVACY

Measure the workflow, not the person

Separate local analysis, fleet telemetry, and team collaboration

PLANE A / FLEET FINDINGS

Person-blind by construction

- ☐ Closed bounded schema
- ☐ Agent and task classes
- ☐ k-gated aggregate cells
- ☐ No person or email field
- ☐ Cannot join to collaboration

PLANE B / COLLABORATION

Identity-bearing on purpose

- ☐ Workspace members and roles
- ☐ Vault indexes and links
- ☐ Task handoff and sharing
- ☐ Tenant is privacy boundary
- ☐ Separate tables and routes

Questions for internal review

- ☐ Which raw transcripts and source paths remain local?
- ☐ Which bounded findings may enter a shared workspace?
- ☐ Who can invite members, mint machine tokens, and revoke access?
- ☐ Which cloud-only or remote sessions are outside current coverage?
- ☐ Can any fleet metric resolve to an individual?

No employee-productivity promise

The baseline examines workflows, coverage, evidence, and cost shape. It does not rank developers or infer productivity from prompts.

05

LOCAL AUDIT

Run the qualification step before buying

A technical owner can start with one command and one workflow

01

Install

Install the MIT-licensed plugin for Claude Code, Codex, or Cursor. The six analyses need no cloud account.

02

Run /qruns

Inspect coverage, scheduled and child runs, terminal state, successful writes, and current local artifact evidence.

03

Review

Choose one workflow whose evidence can change an operating, rollout, or governance decision.

Bring this to the review

- ☐ The workflow and expected output
- ☐ Approximate run frequency and machines
- ☐ The last silent failure or cost surprise
- ☐ Which environments are local, remote, or cloud-only
- ☐ The decision and date the evidence should support
- ☐ The privacy and approval constraints

THE AUDIT QUALIFIES THE PAID WORK

If the report contains no decision-quality signal, there is no baseline proposal. If it does, scope one workflow for a fixed two-week engagement.

```
$ /qruns                                SESSION-VIZ.COM/#START
```

06

DECISION

Agent Fleet Baseline readiness sheet

Use this fillable page to decide whether a two-week baseline is warranted

WORKFLOW

- [ ] One important workflow is named
- [ ] Expected evidence is explicit
- [ ] A technical owner is available
- [ ] A decision date exists

COVERAGE

- [ ] Harness is supported
- [ ] Relevant transcripts are local
- [ ] Remote gaps are documented
- [ ] One machine can run the audit

BOUNDARY

- [ ] Local and shared data are agreed
- [ ] No individual ranking is requested
- [ ] Unknowns may remain unknown
- [ ] Exclusions are understood

OUTCOME

- [ ] Audit review is scheduled
- [ ] One intervention is possible
- [ ] Before/after may be observed
- [ ] Approver and budget path are known

DESIGN-PARTNER BASELINE

Two weeks. One workflow. One operating decision.

EUR 1,500-3,000 after confirming machines and scope. Request a 30-minute fit check at session-viz.com/fleet-baseline/.