

From the episode “How do you make a podcast with AI?”

What the frontier labs advise for working with AI agents

Nine findings from Anthropic, OpenAI and Google Cloud on building and testing agents, and what to do about each.

01 Start with one agent. Several use 3 to 10 times more tokens (the units you pay for).

Anthropic has seen teams spend months on multi-agent systems, then match them by prompting one agent better. Several agents buy thoroughness, not speed: they often take longer overall. Add one only when the first is overloaded with context, parts can run in parallel, or a task needs its own tools.

Anthropic, Jan 2026 • OpenAI, Apr 2025 • Google Cloud, updated May 2026

02 Split work by what each agent needs to know, not by job title.

In an Anthropic experiment, agents split by role (planner, implementer, tester, reviewer) “spent more tokens on coordination than on actual work.” Planning, doing and testing one piece of work share too much context to split. One split that works: a checker that sees only the finished result (item 5).

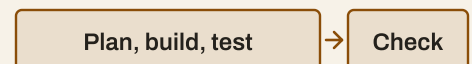
Anthropic, Jan 2026

SPLIT BY JOB TITLE



More tokens on coordination than on the work itself.

SPLIT BY WHAT EACH NEEDS TO KNOW



The checker sees only the finished result.

03 Brief it like a brilliant new employee, and give your reasons.

Anthropic's example: “NEVER use ellipses” works less well than explaining that a speech engine will read the text aloud and cannot pronounce them, because the model generalizes from the reason. If a colleague with no background would be confused, so will the model. Your existing procedures make good first instructions.

Anthropic, read Oct 2026 • OpenAI, Apr 2025

04 Give it a check it can run, and ask for evidence.

Without one, an agent stops when the work looks done, and you become the checker. Telling the model “**Before reporting progress, audit each claim against a tool result from this session**” nearly eliminated fabricated status reports in Anthropic's testing, even on tasks designed to provoke them.

Anthropic, read Oct 2026

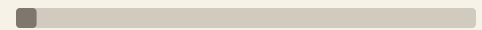
05 Don't let the maker grade its own work.

Agents judging their own work tend to praise it, even when it is plainly mediocre. A separate reviewer is easier to make skeptical, but check it: Anthropic's early reviewer found real problems, then approved the work anyway. Pay for one only where the model is not yet reliable alone.

Anthropic, Mar 2026 · Google Cloud, updated May 2026

SAME REQUEST, TWO SETUPS

One agent



\$9 · 20 min · the game didn't work

Planner, builder, separate reviewer



\$200 · 6 hr · the game played

06 Before a long job, write down what done means.

Anthropic's long-running agents tried to do everything at once, then declared the job finished early. What worked: a written list of everything “done” includes, each item marked not done until tested; one item at a time; each new session starting by reading the progress notes and running a quick test.

Anthropic, Nov 2025

07 If an agent runs on a schedule, a source it can't reach looks like a quiet day.

An expired login looks like “nothing new,” so flag failed reads. Read from where the last run stopped, not “the last 24 hours,” and re-check before posting: one stale item “costs more trust than ten missing items.” Give read-only access where it only reads: others' text can act as instructions.

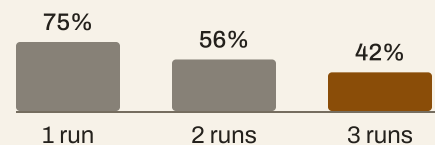
Anthropic, Oct 2026

08 Build a small test set from real failures.

“20-50 simple tasks drawn from real failures is a great start,” Anthropic writes. Write each so two experts would agree on pass or fail. Run each more than once: at 75% success per run, an agent passes three runs in a row only about 42% of the time.

Anthropic, Jan 2026

PASSES EVERY RUN, AT 75% PER RUN



0.75 per run, multiplied:
Anthropic's formula.

09 When the model changes, test again and cut old instructions.

Anthropic found old prompts on a new model cost 36% more per ticket at equal accuracy. Cleaned up, the same prompts cost 14% less than before the cleanup and got 97% of tickets right, up from 92%. Removing “verify twice” cut cost by a third: telling it to double-check is not a check it runs.

Anthropic, read Oct 2026

SOURCES, BY ITEM

Undated pages carry the day we read them, 12 Oct 2026. Links: written version on workhacker.com.

1, 2 **Building multi-agent systems.** Anthropic, Jan 2026

1, 3 **A practical guide to building agents.** OpenAI, Apr 2025

1, 5 **Choose a design pattern for your agentic AI system.** Google Cloud, updated May 2026

3 **Prompting best practices.** Anthropic, read Oct 2026

4 **Best practices for Claude Code.** Anthropic, read Oct 2026

4 **Prompting Claude Fable 5.** Anthropic, read Oct 2026

5 **Harness design for long-running application development.** Anthropic, Mar 2026

6 **Effective harnesses for long-running agents.** Anthropic, Nov 2025

7 **Building effective agent automations.** Anthropic, Oct 2026

8 **Demystifying evals for AI agents.** Anthropic, Jan 2026

9 **Optimizing for cost and intelligence.** Anthropic, read Oct 2026