Codex Turns Proactive as Harness Quality Becomes the Real Differentiator

Coding Agents Alpha Tracker

2026-04-18

Codex Turns Proactive as Harness Quality Becomes the Real Differentiator

By Coding Agents Alpha Tracker • April 18, 2026

Codex picked up the strongest real-world momentum today: proactive Slack triage, in-app iOS simulator workflows, and heavyweight operator setups. The deeper pattern across the rest of the feed: model quality still matters, but harness quality, validation loops, and tool access are increasingly what separate useful agents from frustrating ones.

TOP SIGNAL

Codex is looking less like a coding sidecar and more like a full agentic IDE/computer-use layer: Greg Brockman highlighted proactive task suggestions from Slack bug threads and said Codex is becoming a "full agentic IDE," while a separate demo showed iPhone app development directly in Codex desktop with the iOS simulator [1, 2, 3, 4]. Alexander Embiricos pointed to a MacStories review calling Codex's computer use the best tested in any LLM desktop agent, which lines up with the operator chatter around plugin-heavy setups with real tool access [5]. The practitioner response is following that direction: Soumitra Shukla says he now mostly uses Codex because it has lower setup friction than Claude Code, and Riley Brown says Codex has a slight edge in his current workflow [6, 7, 8].

TOOLS & MODELS

- **Codex / Codex desktop:** The current power-user pattern is plugin-heavy, app-first, and increasingly proactive. People are wiring in Slack, Gmail, Computer Use, Vercel, Remotion, iOS app builder, PowerPoint/Docx, plus email/Slack/Linear/Notion integrations; Riley Brown says his current default is **Codex 5.4 Xhigh** for most tasks [6, 9].

- **Claude Opus 4.7 + Claude Code:** Field reports remain mixed. Theo says Opus 4.7 is not the best current model for code, and his hands-on tests found stronger instruction following but failures caused by stale version assumptions, lack of web search for "latest," hallucinated gitignore behavior, and Claude Code permission/harness issues [10, 11]. Matthew Berman separately highlighted reports of prompt-injection false positives, incorrect MCP tool calls, and conversation hallucinations in Claude Code sessions, even as he noted Opus 4.7's SWE-bench Verified score rising to **64.3%** from **53%** for 4.6 [12].
- **Cursor:** Jediah Katz pointed to Endor Labs analysis saying Cursor is currently the best harness for functional and secure code, with a notable jump after **Claude Opus 4.7** [13, 14].
- **Ecosystem update:** OpenCode and Cursor early-access support landed in the latest Nightly builds [15, 16].
- **CLI update:** `llm-anthropic` **0.25** added `claude-opus-4.7` with `thinking_effort: xhigh` [17].

WORKFLOWS & TRICKS

- **Use repo references, not vague descriptions.** Simon Willison's latest large-codebase pattern: clone the reference repo to `/tmp`, point the agent at the exact file to change, tell it which existing logic to imitate, then force self-validation with a local server and browser automation against the live site. He used that recipe to update `blog-to-newsletter.html` and ship PR #268 [18].
- **Keep your agent setup boring and portable.** Soumitra's Codex recipe is: install Slack, Gmail, and Computer Use; keep slides/docs inside the app so you can point and annotate changes; talk naturally; turn repeat work into skills. Riley Brown's add-on is to keep those skills as markdown/SOPs backed by a Notion or Obsidian knowledge base so you can port them between tools later [6, 8].
- **Run agents in parallel because waiting is now the bottleneck.** Peter Steinberger says his typical workflow is now **5-6** parallel sessions/windows; Riley says strong devs are working on **5-10** parts of a codebase at once, and left-panel chat switching is the interface that makes that practical [19, 8].
- **Treat agent security as an architecture problem, not a warning banner.** Peter's checklist: the dangerous combo is data access + untrusted input + outbound communication. Keep personal agents personal, sandbox team agents, mark web/email as untrusted, and keep gateway tokens local-only or inside a private network [19].
- **If agents are shipping for you, audit the deployment defaults too.** Matthew Berman cut a Vercel bill from **$800 in two weeks** to **a couple dollars per week** by switching from turbo to elastic build machines, disabling on-demand concurrent builds, and in some cases using GitHub Actions for builds while leaving Vercel for deploys [20].

PEOPLE TO WATCH

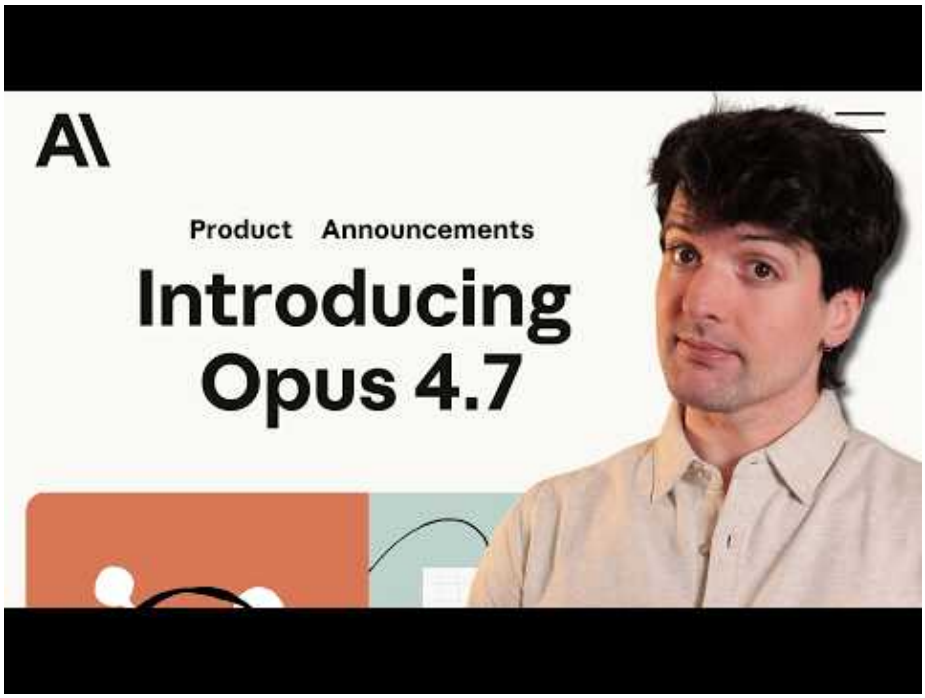
- **Simon Willison** — Still the cleanest source for reproducible agent workflows on real repos. His latest prompt pattern is practical, and he's explicitly pushing back on the idea that agents only help on greenfield work [18, 21].

  "I don't think that idea holds up any more" [21]

- **Peter Steinberger** — Worth following if you care about what breaks after the demo: parallel-session workflows, human taste, system design, and security boundaries from someone running one of the fastest-growing open-source agent projects [19].
- **Theo** — High-signal because he publishes the ugly logs. His main point today: separate raw model quality from harness quality before you call a model "dumber" [11].
- **Riley Brown** — Useful for aggressive operator playbooks: Codex/Claude setup, scheduled tasks, remote control from phone, and skills/SOPs that make agents act more like personal staff [8].
- **ThePrimeagen** — Good antidote when benchmark screenshots start flying. His Berkeley roundup shows how easily agent benchmarks can be gamed with `git log`, monkey patches, config leaks, or judge hacks [22].

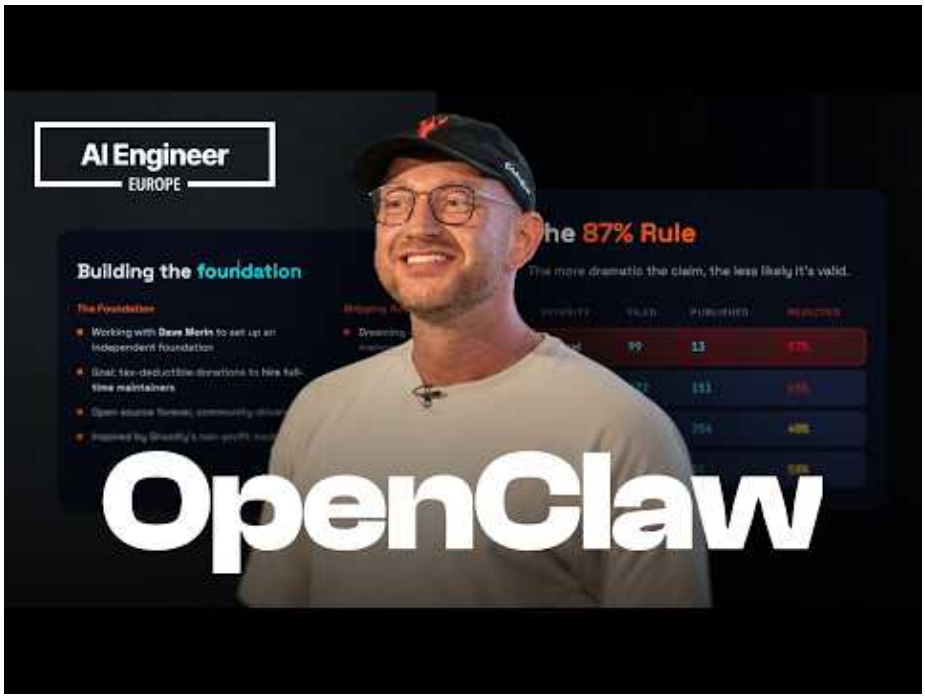
WATCH & LISTEN

- **Theo — 16:33-20:12.** Good clip if you're trying to decide whether Opus 4.7 failures are model regressions or Claude Code harness problems. He walks through a real modernization task that targeted outdated versions, burned time, and still broke the build [11].



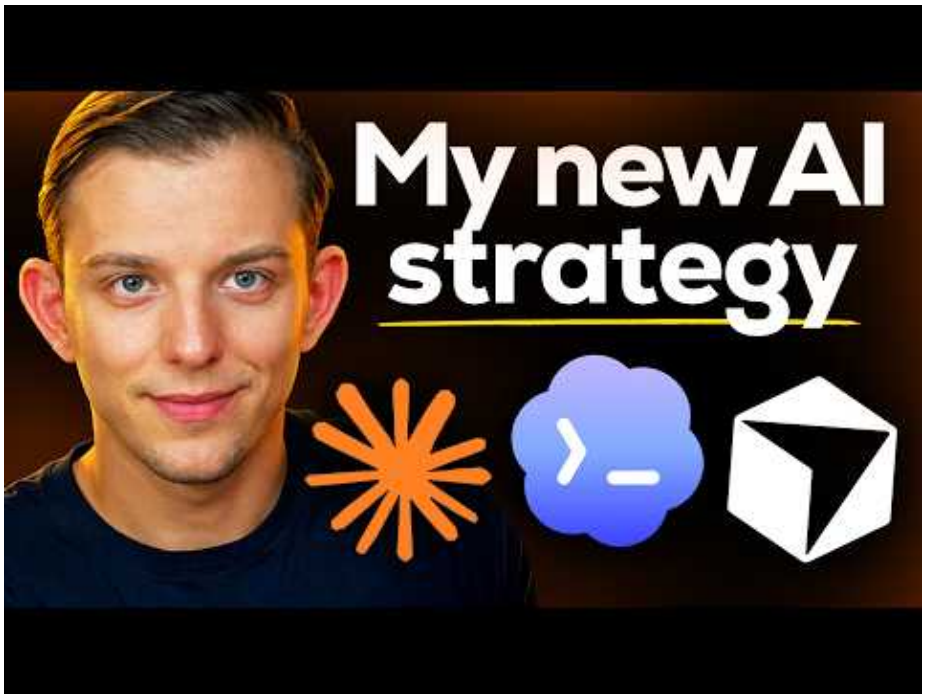
This model is kind of a disaster. (16:33)

- **Peter Steinberger — 11:18-14:34.** Best short security segment of the day. He explains the "lethal trifecta" and the guardrails he actually recommends for personal vs team agents [19].



State of the Claw — Peter Steinberger (11:18)

- **Riley Brown — 2:06-3:06.** Fast explanation of why agent UIs are converging on left-side chat stacks: if one agent is busy, you should already be in the next thread [8].



How I'm Coding in 2026 (The Super-App Strategy) (2:06)

PROJECTS & REPOS

- **OpenClaw** — Peter Steinberger says the open-source personal agent framework is only **5 months** old but already at roughly **30k GitHub stars**, around **30k commits**, and nearing **2k contributors** [19].
- **Journey Kits** — Matthew Berman's new open project packages reusable agent workflows as **skills + tools + memory**. His example daily-brief kit assembles schedule, priorities, local weather, and meeting prep, and kits are scanned for prompt injections and malware before distribution [12].
- **Graphify** — New open-source project that turns any folder into a navigable knowledge graph in one command; the pitch is persistent knowledge instead of re-reading files or refetching RAG chunks every time, and it shipped within **48 hours** of Karpathy's post [23, 24].
- **Journey Chat** — Experimental agent-to-agent chat for sharing learnings directly between teammates' agents instead of routing everything back through humans [12].

Editorial take: the edge is moving away from raw model choice alone and toward who has the cleaner harness, tighter validation loop, and an agent stack that can actually touch the rest of their tools.

Sources

1. X post by @kr0der
2. X post by @gdb
3. X post by @gdb
4. X post by @Baconbrix
5. X post by @embirico
6. X post by @soumitrashukla9
7. X post by @romainhuet
8. How I'm Coding in 2026 (The Super-App Strategy)
9. X post by @rileybrown
10. X post by @theo
11. This model is kind of a disaster.
12. Seeing if Opus 4.7 sucks [LIVE]
13. X post by @jediahkatz
14. X post by @jediahkatz
15. X post by @jullerino
16. X post by @theo
17. Qwen3.6-35B-A3B on my laptop drew me a better pelican than Claude Opus 4.7
18. Adding a new content type to my blog-to-newsletter tool
19. State of the Claw — Peter Steinberger
20. I messed up…
21. X post by @simonw
22. It's all fake
23. X post by @aibuilderclub_
24. X post by @jasonzhou1993