Per-Commit Codex Loops, DeepSeek V4 Flash, and Script-Writing Search Agents

Coding Agents Alpha Tracker

2026-04-29

Per-Commit Codex Loops, DeepSeek V4 Flash, and Script-Writing Search Agents

By Coding Agents Alpha Tracker • April 29, 2026

Peter Steinberger's live setup puts Codex on every commit, with auto-fix PRs and review loops catching regressions fast. Also here: DeepSeek V4's price/performance story, local Flash inference on 128GB Macs, Sourcegraph's script-executing Deep Search, and a few operator-grade workflow lessons.

TOP SIGNAL

The biggest practical shift today is **agentic CI/CD**. Peter Steinberger says Codex now reviews every landed commit, spawns a fresh Codex to open a fix PR when it finds a bug, then hands that PR to a review/fix loop that can run up to five times; in a separate per-commit-to-main setup, it found one of his own regressions within 10 minutes. [1, 2]

Timeless takeaway: put the agent on the commit path, keep the loop bounded, and use PRs as the handoff + audit boundary. [1, 2]

TOOLS & MODELS

- **DeepSeek V4 Flash** — Salvatore Sanfilippo says Flash is the real local-agent story in the V4 release: he implemented **2-bit asymmetric quantization**, runs it on a **128GB MacBook**, and says **tool calling works perfectly**. He compares it to recent Sonnet-level performance, while being more cautious than frontier-model claims. [3]
- **Flash vs. Pro** — Same source argues **V4 Flash**, not Pro, is ready now for local inference, and says Pro's training is not finished yet. He also says Flash beats **Kimi 2.6** at roughly **1/3 the size** and has more disciplined thinking behavior than Qwen 3.6. [3]

- **Local throughput reality** — On his MacBook, Sanfilippo reports about **120-130 TPS prefill** and **21+ TPS generation**, with the warning that **prefill is the real bottleneck** for coding agents. [3]
- **Sourcegraph Deep Search** — now writes and executes scripts to analyze codebases, then feeds results back into the agent. Sourcegraph frames this as **custom tools on demand**; example query: "Top Files in VS Code Bugfix Commits (Last 6 Months) and Contributors." changelog [4, 5]
- **Cursor / EndorLabs benchmark signal** — Jediah Katz shared EndorLabs' latest correctness-and-security benchmark, which he says had **Cursor's optimized harness** on top. Useful if you're tracking harness quality, not just model choice. [6, 7]
- **Codex endurance** — Tibo says that with some small tweaks, **Codex can work for days on hard tasks**, and that changes are coming to make this easier to use. [8]

WORKFLOWS & TRICKS

- **Per-commit auto-fix loop**
  1. Run Codex on every landed or main-branch commit.
  2. If it finds a regression or security issue, spawn a new Codex instance to open a fix PR.
  3. Hand that PR to a review agent.
  4. If review finds problems, spawn another fix agent and loop again — Steinberger caps it at **5 passes**.
  5. Use the PR as the audit trail. Example PR and example commit record [1, 2]
- **Script-escape pattern for repo analysis** — When the native agent loop gets stuck, let the agent write and run a one-off script, then feed the result back into the loop. That's the core pattern in Sourcegraph's Deep Search update, and it generalizes to churn analysis, migration prep, and codebase archaeology. [4]
- **Measure the latency that actually hurts** — For local coding agents, Sanfilippo says **prefill**, not generation, is the real constraint. If you're evaluating quantization or model swaps, benchmark the part that blocks long-context coding work. [3]
- **Expect API exhaustion once agents scale** — Steinberger says agents can hit GitHub rate limits even after a move to Enterprise. In his sessions, Codex worked around GitHub limits via the browser, typed into a comment box to close an issue, and opened Cloudflare to create a new API key when permissions were missing; he also plans to test ghx. [9, 10, 11, 12]
- **Claude Code config check** — AI Builder Club warns that many setup guides still recommend deprecated `npm install`, and Jason Zhou called out the overlooked `.claude/rules/` directory. Small detail, real leverage. [13, 14]

PEOPLE TO WATCH

- **Peter Steinberger** — strongest operator signal in today's notes: per-commit Codex reviewers, live fix PRs, browser fallbacks, and the ugly reality of GitHub API saturation. [1, 9, 10]
- **Salvatore Sanfilippo** — brings the local-inference details most model chatter skips: quantization method, RAM target, throughput numbers, and a clear argument that **Flash**, not Pro, is the local-agent story right now. [3]
- **Daniel Neal Adler / Sourcegraph** — high signal if you care about agents that inspect large codebases, because he's shipping the "write code to understand code" pattern into product. [4]
- **Jason Zhou** — useful for catching low-drama but high-leverage setup details like deprecated Claude Code install paths and hidden config surfaces. [13, 14]
- **Tibo** — short post, strong implication: long-running Codex sessions are getting easier, which matters if your hardest tasks aren't one-shot edits. [8]

WATCH & LISTEN

- **0:02-2:32 — Why DeepSeek V4 Flash matters more than Pro for local agents.** Sanfilippo walks through the practical case: 2-bit asymmetric quantization, 128GB Mac target, working tool calls, and why Flash is the interesting part of the release for local inference. [3]

La perla di casa DeepSeek è il modello Flash, non il Pro, almeno per ora

(0:01)

- **21:02-23:58 — Codex using the browser as a test harness.** Riley Brown shows Codex turning an HTML file into an app and then validating buttons, navigation, and quiz flows by controlling the browser itself. [15]

Learn 95% of Codex in 30 minutes (21:02)

PROJECTS & REPOS

- **openclaw/openclaw PR example + clawsweeper commit record** — best repo-level signal today because it's live evidence of agentic CI, not a concept deck. One link shows the fix-PR loop; the other shows a regression caught almost immediately after launch. [1, 2]
- **ghx** — niche, but relevant if your bottleneck is GitHub API exhaustion rather than model quality; Steinberger singled it out while troubleshooting agent-heavy workflows. [9]

Editorial take: today's durable edge is orchestration — commit hooks, bounded review/fix loops, script escapes, and browser fallbacks matter as much as the base model. [1, 4, 10]

Sources

1. X post by @steipete
2. X post by @steipete

3. La perla di casa DeepSeek è il modello Flash, non il Pro, almeno per ora
4. X post by @DanielNealAdler
5. X post by @Sourcegraph
6. X post by @jediahkatz
7. X post by @jediahkatz
8. X post by @thsottiaux
9. X post by @steipete
10. X post by @steipete
11. X post by @steipete
12. X post by @steipete
13. X post by @aibuilderclub_
14. X post by @jasonzhou1993
15. Learn 95% of Codex in 30 minutes