

# Simon Willison: agent-built services need hard budget caps by default

Coding Agents Alpha Tracker

2026-10-04

## Simon Willison: agent-built services need hard budget caps by default

*By Coding Agents Alpha Tracker • October 4, 2026*

Willison argues that hard spend limits should be the default for anything an agent deploys. Also: harness choice moves benchmark scores sharply, T3 Code can queue prompts until usage limits reset, Claude Code adds a “You should know” mod, and the Codex updates from DevDay.

### Put hard spend caps on anything your agents deploy

Simon Willison argues that coding agents, and personal agents (which he calls “coding agents wrapped in a less threatening UI”), make it easy to spin up code that costs money: paid API calls, hosted apps, and storage or compute that bills as it grows [1]. A warning email is not enough. Usage-based services need **hard** caps that cut the service off and return errors. Removing the cap should be an explicit opt-in [1]. His reasoning: most people would rather see errors than a surprise bill of \$10,000 or more [1].

Options you can use now: - **AWS** announced monthly project spend limits on Sept 16. A project that hits its limit is paused for the rest of the month. AWS’s docs say the feature is still going out to a limited number of customers [1]. - **Google Cloud** launched Spend Caps in July. They set a monthly cap on specific services within a project [1].

Willison also wants agents to lean toward recommending providers with hard caps, and to warn inexperienced builders before they deploy to uncapped services [1]. You can do this today by adding that rule to your agent instructions.

## The harness can matter as much as the model

Two results summarized by AINews point the same way. Hugging Face reports that the same model weights scored 62% in one harness and 33% in another. Its multi-harness RL setup uses a proxy that records token IDs and logprobs without changing the harnesses. It lifted LFM2.5-2.6B from 42% to 54% across four harnesses with 31% fewer tool calls, and the trainer, data and seven models are open [2]. Separately, Meta Superintelligence Labs reports that a dedicated controller raised GPT-5.5 on ProgramBench from 63.7% to 71.5%, with the same workers and budget, against 58.0% for Codex [2]. In practice, a model comparison means little unless you know which harness ran it.

## T3 Code: queue prompts until limits reset, see spend across tools

- **Queued messages:** T3 Code Nightly now lets you queue messages that fire when your usage limits reset [3].
- **Usage view:** Theo improved the view so you can see where your “spend” goes and how models behave on your own data. It counts all Claude Code and Codex usage on your machines, not just usage inside T3 Code [4].
- **Model mix:** Opus 5.5 is the first model to pass 50% of T3 Code traffic; “literally half of all prompts go to Opus” [5].
- **Orchestrator V2** is in the latest Nightly, and Theo is asking for reports on what broke or confused people [6]. The mobile app only works through the TestFlight/beta build (install docs) [7].
- **Users:** T3 Code passed 400K users and gained about 20K more within a day [2, 8].

## Claude Code mods in use

Anthropic added a built-in plugin, “You should Know.” It scans Claude’s output for important information you might miss. Enable it with `/plugin enable cc-plugin-you-should-know@builtin` [9]. AINews describes it as spinning off a side agent, and describes mods as plugins with middleware-like hooks into Claude Code [2]. A community example: @shawnbuilt built a custom Jev memory harness as a mod and posted the full prompt to rebuild it [10].

On a related habit, swyx says: “always run some kind of skill review/cutter after every model-created skill” [11].

## Codex: DevDay features and a product refocus

Riley Brown’s DevDay walkthrough covers the features most relevant to Codex users: - **Codex CLI:** easier voice input, a new agents view and better prompt editing [12]. - **Codex Cloud:** you set up an environment once (signed into Clerk, Convex, Vercel, GitHub and so on), then reuse it for tasks run from your phone, with no local machine left on [12]. - **Code review:** PRs can be reviewed

from a side panel. You may need to pin “code review” there first [12]. - **GPT-6.1 Sol:** OpenAI pitches it as a lower-cost model for coding and computer use. Brown quotes \$2 in / \$10 out per million tokens, against \$10 / \$50 for GPT-6 Astra [12]. - **Ultra-fast mode:** about 8x faster output at 6x the usage, and only on the \$500/month plan [12].

OpenAI’s Tibo says the Codex team is “locking in”: the only work now underway is simplification, efficiency to allow more usage, groundbreaking features and new models, because feedback says users want things simpler [13].

## Smaller items

- **A dot clearing an inbox:** Tibo reached inbox zero by giving his dot the goal. It deleted unneeded categories in validated batches, labeled email by type of work, and walked him through replies while it looked up context in the background [14]. Simon Willison, by contrast, isn’t sure when to use a dot over ChatGPT, because his dot seems built around a single conversation while he prefers managing context across threads [15].
- **Cheap computer-use targeting:** ThePrimeagen uses Cloudflare’s decision API for computer use. He raises its probabilities to chosen powers, then applies “probabilistic centering” so a command like “Click the monitor icon in the menu bar” lands on the right icon [16].
- **Pi 1.0** now includes native MCP support in Codemod by default, plus deferred tool loading, Anthropic cache warming and mid-conversation system messages [2].
- **Apple:** Apple plans new Mac privacy controls that warn about granting broad data access to third-party software, including AI agents [17]. DHH expects this to make macOS harder to use productively with agents [18].

---

## Sources

1. We’re going to need default hard budget caps on pretty much everything
2. [AINews] not much happened today
3. X post by @theo
4. X post by @theo
5. X post by @theo
6. X post by @theo
7. X post by @theo
8. X post by @theo
9. X post by @ClaudeDevs
10. X post by @shawnbuilds
11. X post by @swyx
12. ChatGPT Just Changed A LOT (Here’s Everything New)
13. X post by @thsottiaux
14. X post by @thsottiaux

15. X post by @simonw
16. X post by @ThePrimeagen
17. X post by @natlungfy
18. X post by @dhh