

Codex's Usage Reset Exposes Unbounded Agent Loops

Coding Agents Alpha Tracker

2026-08-30

Codex's Usage Reset Exposes Unbounded Agent Loops

By Coding Agents Alpha Tracker • August 30, 2026

A Codex audit found quota-burning failures in compaction, goals, automations, subagents, and MCP; the practical response is regression-testing agent control paths before unattended runs.

TOP SIGNAL

Codex's usage reset exposes control-plane bugs in autonomous coding agents. @thsottiaux says paid Codex and ChatGPT Work users are being reset after thousands of reports, with expected usage 10%–50% further depending on how Codex is used; the audit found retained-image compaction loops, `/goal` overruns or retries that consumed 15%–70% of a weekly allowance, over-frequent automations, unexpected helper escalation, and duplicated or truncated MCP results. [1]

Treat those as a regression suite: in a disposable repo, replay an image-heavy compaction, a `/goal` with a deliberate stop, a Stop-hook interaction, a custom schedule, and a large MCP result; log token deltas, helper selection, and whether the run terminates. This test plan is my recommendation, based on the failure modes the update says were fixed. [1]

TRY THIS

- **Turn idea dumps into a pull queue.** ThePrimeTime's Linear MCP workflow is simple: finish a long brainstorm, have the agent turn the ideas into tickets, then—when you do not know what to work on—run: **“Hey, Agent, go to Linear and tell me what I need to do next.”** Resource: Linear Terminal. [2]

- **Make a side-effecting agent draft-only by construction.** Kent C. Dodds’s Kody pattern starts with the explicit grant “**Create a draft reply. Never send.**” Put that intent in the README and export docs; request `gmail.compose` plus `gmail.googleapis.com`, add `gmail.readonly` only if inbox reading is required, call `users.drafts.create` rather than a send endpoint, then publish a thin drafts-only package and lock it with `changes: { locked: true }`. Gmail’s OAuth scope can still technically send, so the package lock—not the token—is the authority boundary; later publishes require the owner’s **Promote this commit** action. [3]
- **Use real file paths for remote context.** In the T3 Code nightly, attach any file; the attachment is exposed to the agent as a real path. Upload a PDF, Markdown file, or MP3 and let the agent operate on that path instead of building a bespoke transfer step. [4]
- **A/B a cheap model with a deterministic smoke test.** Matthew Berman’s GLM workflow is replicable: create a Z.AI or OpenRouter API key, connect it to OpenCode—or another client supporting an OpenAI-compatible endpoint—and avoid sensitive data when using the China-served Z.AI endpoint. His baseline task was a Rubik’s Cube simulation; he checked scrambling, solving, cube size, colors, speed, camera, lighting, and materials. [5]

WHAT SHIPPED

- **GLM 5.3 Flash (Z.AI) is the main model-level release signal.** Berman presents it as an open-weights mixture-of-experts model with 320B total parameters and 18B active, reporting Terminal Bench 84.3 and DeepSuite 63.4; he also cites a 1M-token context, 131K maximum output, and maximum reasoning by default. [5] His cost estimate is about **\$0.09 per intelligence-index task**, but the model uses about **47,000 output tokens** on average versus roughly 20,000 for Luna; he still calls Luna the better pure cost/quality trade-off, while GLM buys open-weight control. [5] His same-prompt demos are not a controlled benchmark: GPT 5.6 Soul had internet access in Codex while OpenCode did not. [5]
- **Claude Code’s capacity schedule is changing on September 14.** ClaudeDevs says the current 50% weekly-limit increase remains in place until then, after which standard limits for Pro, Max, Team, and seat-based Enterprise plans will be permanently 25% higher than the prior standard; the team characterizes that as a 17% reduction from today’s promotional level. [6, 7] Treat September 14 as the budget boundary for long-running agent work.
- **T3 Code’s adoption signal is now substantial.** Theo reports more than **250,000 users** and **70,000 weekly active users** nine days after an earlier update, and says Linux became the most popular platform as of

August 23. [8, 9] Omarchy 4.1 is slated to ship with T3 Code, with DHH describing setup as taking only a few seconds. [10]

- **Provider exposure needs task-level accounting.** In the Cursor discussion, @thsottiaux cautions that the cited 5% model-traffic share is not a proxy for revenue or value because token-efficient frontier models consume fewer tokens; a current-period video separately quotes Harrison Chase arguing that a harness owned by no model lab is the only architecture that works across every model. [11, 12]

GO DEEPER

- **Matthew Berman — “Cancel your subscriptions, Ox-Alpha is here! (GLM 5.3 Flash)”:** Watch the API hookup, Rubik’s Cube smoke test, and same-prompt comparison. The useful lesson is how to combine a deterministic artifact test with cost, token-efficiency, and tool-access caveats rather than trusting a leaderboard claim alone. [5]



Cancel your subscriptions, Ox-Alpha is here! (GLM 5.3 Flash) (11:19)

- **OpenAI Sets a Date to Cut Cursor Off:** Skip the ownership drama and watch the Amazon/Kiro workflow section. The video reports that, across roughly 50 Amazon Stores teams, merely adding AI produced under 3× prior deployment velocity while teams that changed how they worked reached 4.5× or better, with documentation, precise errors, compiler feedback, and locally mocked services offered as the practical levers. The figures are self-reported

and deployment velocity is not the same as value shipped. [12]



OpenAI Sets a Date to Cut Cursor Off (10:20)

- **Study Kody’s locked Gmail drafts guide:** It is a compact example of separating authentication from declared authority: a published export creates a draft and returns its ID, while the human reviews and sends; the lock prevents this package’s jobs and exports from silently becoming a sender. [3]

Editorial take: The highest-alpha agent work today is not adding autonomy; it is making autonomy observable, interruptible, and narrowly authorized—usage accounting, deterministic smoke tests, and locked side effects are becoming first-class features. [1, 3]

Sources

1. X post by @thsottiaux
2. 1968 Predicted Every Modern Tech Problem | TheStandup
3. Gmail drafts without send — lock what Google cannot scope
4. X post by @theo
5. Cancel your subscriptions, Ox-Alpha is here! (GLM 5.3 Flash)
6. X post by @ClaudeDevs
7. X post by @ClaudeDevs
8. X post by @theo
9. X post by @theo

10. X post by @dhh
11. X post by @thsottiaux
12. OpenAI Sets a Date to Cut Cursor Off