GPT-5.4 rolls out broadly as coding agents, hybrid open models, and interpretability funding accelerate

AI News Digest

2026-03-06

GPT-5.4 rolls out broadly as coding agents, hybrid open models, and interpretability funding accelerate

By AI News Digest • March 6, 2026

OpenAI rolled out GPT-5.4 (Thinking + Pro) across ChatGPT, the API, and Codex—highlighting steering mid-response, 1M-token context, and native computer use—alongside new safety research on chain-of-thought controllability. The digest also covers Cursor's cloud agents workflow, Perplexity's multi-model "Model Council," AllenAI's open Olmo Hybrid architecture release, Goodfire's $150M fundraise, and fresh signals of agents moving into enterprise operations.

OpenAI launches GPT-5.4 (Thinking + Pro) across ChatGPT, API, and Codex

GPT-5.4 roll-out + headline capabilities

OpenAI announced **GPT-5.4** is available now in the **API** and **Codex**, with a **gradual rollout in ChatGPT** starting today [1, 2, 3]. OpenAI frames GPT-5.4 as combining advances in **reasoning, coding, and agentic workflows** into one frontier model [3].

Notable feature claims include:

- **Steering mid-response** (interrupt the model and adjust direction) [4, 1]
- **1M tokens of context** [1]
- Better performance on **knowledge work** and **web search**, plus **native computer use** capabilities [1]
- "Most factual and efficient" (OpenAI claims **fewer tokens** and **faster speed**) [4]

Steering availability: OpenAI says steering is available **this week on Android and web**, with **iOS "coming soon"** [4].

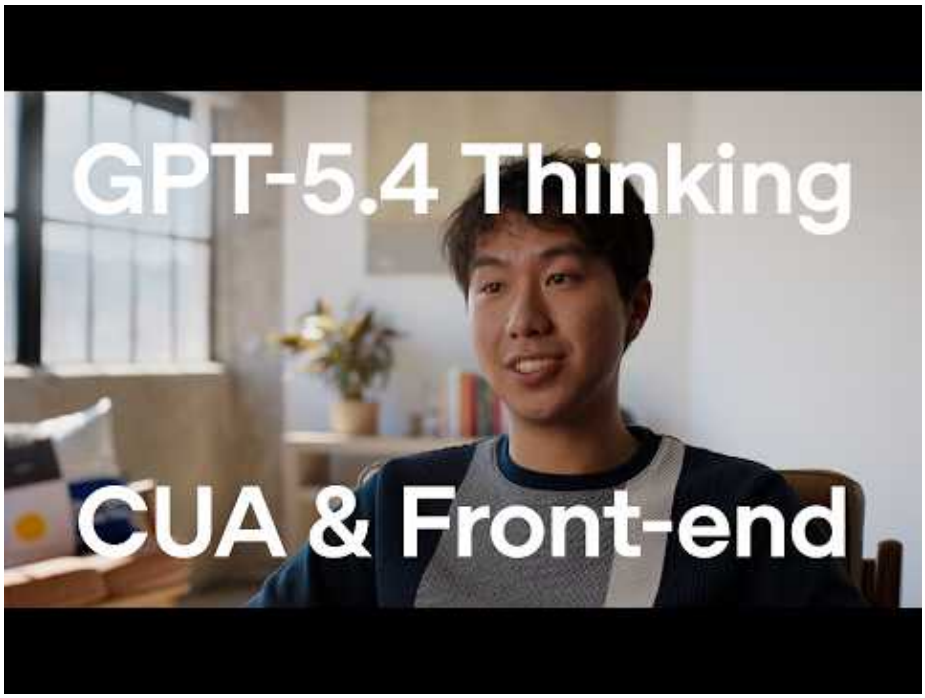
Details: http://openai.com/index/introducing-gpt-5-4/ [2]

Codex updates: /fast mode + GPT-5.4 improvements for tool use and frontend work

Sam Altman also highlighted that **GPT-5.4 is introduced in Codex** with improved **token efficiency** and better **tool calling**, **computer use**, and **frontend development** [5]. OpenAI is also introducing **/fast** for a faster version of Codex [5, 6].

Demo: computer use + frontend workflows (with self-testing efficiency claims)

In OpenAI's GPT-5.4 Thinking demo, the model uses **CUA (computer use)** to test its own work (e.g., interacting with a **3D chess Electron app**) [7]. The presenter claims token use has **dropped by 2/3** in some self-testing cases [7].



Computer Use & Frontend UI with GPT-5.4 Thinking (0:41)

Windows: Codex app + "Windows-native agent sandbox"

OpenAI says the **Codex app is now live on Windows** [8, 9]. The app can run **natively** or in **WSL** with integrated terminals (PowerShell, Command

Prompt, Git Bash, or WSL) [9].

OpenAI also described a **Windows-native agent sandbox** with OS-level controls that block filesystem writes outside a working folder and prevent outbound network access unless explicitly approved [9].

Safety research: Chain-of-Thought (CoT) controllability eval

OpenAI published a new **evaluation suite** and **research paper** on **Chain-of-Thought (CoT) Controllability** [10]. The company reports that **GPT-5.4 Thinking shows low ability to obscure its reasoning**, suggesting CoT monitoring remains a useful safety tool [10].

Paper: https://openai.com/index/reasoning-models-chain-of-thought-controllability/ [10]

Early tester feedback (including weaknesses flagged)

One tester wrote that after a week of testing, GPT-5.4 felt like "the best model in the world" and reduced their reliance on Pro modes [11]. The same thread praised coding reliability in Codex [11] and speed improvements from using fewer reasoning tokens [11].

That tester also listed weaknesses: "frontend taste" lagging competitors, missing obvious real-world context in planning, and stopping short before finishing tasks in OpenClaw [11]. Sam Altman replied: "We will be able to fix these three things!" [12].

Coding agents: Cursor's cloud agents push toward test-and-video workflows

Cursor's "cloud agents" are described as having surpassed tab-autocomplete usage internally, reinforcing the claim that "the IDE is Dead" [13]. In this model, agents do more end-to-end work and return artifacts that are easier to review than raw diffs.

Key product mechanics highlighted:

- **Automatic testing** of changes before PR submission (with calibrated prompting and a `/no test` override) [13]
- **Demo videos** as an entry point for review, plus Storybook-style galleries [13]
- **Remote VM access** (VNC) for live interaction and iteration [13]
- A `/repro` workflow for bug reproduction + fix verification with before/after videos [13]

The same discussion frames a near-term "big unlock" as widening throughput via **parallel agents** and **subagents** for context management and long-running threads [13].

Multi-model orchestration: Perplexity adds "Model Council" to Perplexity Computer

Perplexity launched **Model Council** inside **Perplexity Computer**, allowing users to run **GPT-5.4**, **Claude Opus 4.6**, and **Gemini 3.1 Pro** simultaneously and select an orchestrator model [14]. Perplexity's positioning: "Three frontier models. One workflow. Best answer wins." [14]

Open models and new architectures: AllenAI releases Olmo Hybrid (7B)

Allen AI released **Olmo Hybrid**, a fully open **7B** model combining transformer and **linear RNN (gated delta net / GDN) layers** in a **3:1 ratio** with full attention [15, 16]. AllenAI and commentary in Interconnects describe it as a strong artifact for studying hybrid architectures, with theory and scaling experiments accompanying the release [17].

Interconnects reports:

- Pretraining gains: about a **2× gain on training efficiency** vs. Olmo 3 dense [17]
- Post-training results: mixed (knowledge wins, reasoning losses vs. dense), but still a strong open model overall [17]
- Practical challenge: OSS tooling and long-context inference issues can negate efficiency gains in practice right now [17]

Resources:

- Paper: https://allenai.org/papers/olmo-hybrid [18]
- HF artifacts: https://huggingface.co/collections/allenai/olmo-hybrid [18]
- Analysis: https://www.interconnects.ai/p/olmo-hybrid-and-future-llm-architectures [19]

Research workflow shift: Karpathy's nanochat gets faster—and agents iterate on it autonomously

Andrej Karpathy reported nanochat can now train a GPT-2-capability model in **2 hours** on a single **8×H100** node (down from ~3 hours a month ago), largely due to switching from FineWeb-edu to **NVIDIA ClimbMix** [20].

He also described **AI agents automatically iterating on nanochat**, making **110 changes** over ~12 hours and improving validation loss from **0.862415 → 0.858039** for a d12 model without increasing wall-clock time (feature branch experimentation + merge when ideas work) [20]. Karpathy later framed the "new meta" benchmark as: *"what is the research org agent code that produces improvements on nanochat the fastest?"* [21].

Interpretability funding + "Intentional Design": Goodfire raises $150M Series B

Mechanistic interpretability startup **Goodfire** announced a **$150M Series B** at a **$1.25B valuation**, less than 2 years after founding [22]. Alongside the raise, the company introduced **Intentional Design**: complementing reverse engineering with an approach focused on shaping the **loss landscape** to influence what models learn and how they generalize [22].

One proof-of-concept described is hallucination reduction using a probe trained to detect hallucinations for both **runtime steering** and **RL reward signals**, with a key training trick: run the probe on a **frozen copy** of the model to reduce incentives/ability to evade the detector during training [22].

Enterprise adoption notes: MUFG + Sakana AI lending agent moves to real-case testing; Microsoft updates Dragon Copilot

Sakana AI and Mitsubishi UFJ Bank (MUFG) advanced their "**AI Lending Expert**" agent system from a ~6-month PoC to a **real-case verification phase**, following their 2025 comprehensive partnership announcement [23, 24].

Microsoft announced "big updates" to **Dragon Copilot** at HIMSS, introducing **Work IQ** to bring the right work context alongside patient data, aiming to reduce admin busywork and let clinicians focus more on patients [25].

Two cautionary notes circulating: benchmarks and moral-reasoning behavior

- **Benchmark noise:** swyx cautioned against a viral claim that Claude Opus 4.6 had its "worst benchmark day," pointing out that the SWE-bench author does not endorse "cheap sample" benchmarks and arguing **30–60× more compute** is needed for statistically meaningful results [26].
- **Moral-reasoning oddities:** Gary Marcus amplified a study thread reporting that GPT answered "yes" to torturing a woman to prevent a nuclear apocalypse but "absolutely not" to harassing a woman in the same scenario—described as a reversal that appeared only when the target was a woman [27]. The thread argues this may reflect mechanical overgeneralization from RLHF rather than reasoning about underlying harms [27].

Sources

1. X post by @sama
2. X post by @OpenAI

3. X post by @OpenAI
4. X post by @OpenAI
5. X post by @ah20im
6. X post by @sama
7. Computer Use & Frontend UI with GPT-5.4 Thinking
8. X post by @sama
9. X post by @ajambrosino
10. X post by @OpenAI
11. X post by @mattshumer_
12. X post by @sama
13. Cursor's Third Era: Cloud Agents
14. X post by @AskPerplexity
15. X post by @allen_ai
16. X post by @natolambert
17. Olmo Hybrid and future LLM architectures
18. X post by @natolambert
19. X post by @natolambert
20. X post by @karpathy
21. X post by @karpathy
22. Don't Fight Backprop: Goodfire's Vision for Intentional Design, w/ Dan Balsam & Tom McGrath
23. X post by @SakanaAILabs
24. X post by @hardmaru
25. X post by @satyanadella
26. X post by @swyx
27. X post by @ValerioCapraro