GPT-5.5 Debuts as Codex Broadens Beyond Coding

AI News Digest

2026-04-24

GPT-5.5 Debuts as Codex Broadens Beyond Coding

By AI News Digest • April 24, 2026

OpenAI dominated the day with GPT-5.5, a wider Codex push, and a clearer inference-at-scale message. DeepSeek and Google DeepMind added two more signals: open models are still pushing on context and cost, while training infrastructure is getting more resilient.

OpenAI turns GPT-5.5 into the day's center of gravity

GPT-5.5 arrives with longer context, lower token use, and near-term API access

OpenAI says GPT-5.5 is a new class of intelligence for real work and agents: it is built to understand complex goals, use tools, check its work, and carry tasks through to completion [1]. It is rolling out now to Plus, Pro, Business, and Enterprise users in ChatGPT and Codex, with GPT-5.5 Pro for Pro, Business, and Enterprise in ChatGPT; API access is planned very soon after security and safeguards work [2, 3].

Quick facts: - **Context window:** 1M tokens in the API [4] - **API pricing:** $5 per 1M input tokens and $30 per 1M output tokens [4] - **Serving profile:** GPT-5.4-like per-token latency, but significantly fewer tokens per task [5, 3]

Artificial Analysis says GPT-5.5 (xhigh) now leads its Intelligence Index and tops GDPval-AA, Terminal-Bench Hard, and APEX-Agents-AA; it also says GPT-5.5 (medium) matches Claude Opus 4.7 (max) at roughly one-quarter the cost. Its notable caveat is reliability: on AA-Omniscience, GPT-5.5 posted the highest accuracy at 57% but an 86% hallucination rate, above Opus 4.7 and Gemini 3.1 Pro Preview [6].

Why it matters: The launch ties raw model performance to token efficiency,

pricing, and product availability, which is the mix needed for heavier agent workflows [5, 4, 1].

Codex moves closer to general computer work

With GPT-5.5, OpenAI says Codex now gets more of the job done across the browser, files, docs, and the computer itself; it can interact with web apps, test flows, click through pages, capture screenshots, and iterate until a task is finished [7]. Greg Brockman said the combination is no longer just for coders but for broader computer work, including spreadsheets and slides, and OpenAI is now rolling Codex out across whole companies after a pilot with NVIDIA [8, 9, 10].

Early users described a step change in autonomy. At Ramp, GPT-5.5 plugged into an internal harness and began discovering how to use databases and telemetry tools without explicit guidance; inside OpenAI, one engineer said it produced a tidal wave of pull requests and cases where the model worked on a single task for more than 40 hours [11, 12]. OpenAI also turned on **auto-review** in Codex, where a guardian agent evaluates higher-risk actions so long tasks can continue with fewer approvals [13, 14].

Why it matters: The noteworthy change is not only a stronger model, but a broader product surface for handing off longer, multi-tool workflows inside real organizations [7, 9, 13].

The subtext: inference is becoming strategy

Sam Altman separately said OpenAI now has to become an AI inference company and praised the inference team for serving GPT-5.5 efficiently [15].

> "To a significant degree, we have to become an AI inference company now." [15]

NVIDIA said GPT-5.5-powered Codex runs on GB200 NVL72 systems, that more than 10,000 NVIDIA employees are already using it, and that those systems deliver 35x lower cost per million tokens and 50x higher token output per second per megawatt than prior-generation platforms [16]. NVIDIA also described the launch as part of a long-running OpenAI partnership that now includes a commitment to deploy more than 10 gigawatts of NVIDIA systems for next-generation AI infrastructure [16].

Why it matters: On the same day as a flagship model launch, both OpenAI and NVIDIA framed the bottleneck as serving capable agents economically and at enterprise scale, not only training the next model [15, 16].

The rest of the field kept moving on openness and infrastructure

DeepSeek open-sources V4 with 1M context

DeepSeek said its V4 preview is live and open-sourced, pitching cost-effective 1M context length. The release includes **V4-Pro** at 1.6T total / 49B active parameters and **V4-Flash** at 284B total / 13B active parameters, with updated API access and public weights and technical report links [17].

Emad Mostaque estimated the final training runs at under $14M for Pro and under $4M for Flash, with total compute across data prep, tuning, and testing around 10x those figures [18].

Why it matters: The open-model push is still advancing on long context and cost claims at the same time, rather than conceding those fronts to closed labs [17, 18].

Google DeepMind shows more failure-tolerant frontier training

Google DeepMind introduced **Decoupled DiLoCo**, a system for training advanced models across multiple data centers without halting when hardware fails [19, 20]. The company says it combines Pathways and DiLoCo, is self-healing during induced failures, trained a 12B Gemma model across four U.S. regions over low-bandwidth networks, and mixed TPUv5p with TPU v6e without slowing training [21, 22, 23].

Jeff Dean said the approach lets (N-1)/N units proceed when one fails, framing it as a continuation of Google's long-running work on large-scale fault-tolerant training [24, 25].

Why it matters: As cluster sizes grow, resilience across regions and hardware types is becoming a research advantage in its own right [23, 24].

Sources

1. X post by @OpenAI
2. X post by @OpenAI
3. X post by @sama
4. X post by @sama
5. X post by @OpenAI
6. X post by @ArtificialAnlys
7. X post by @OpenAIDevs
8. X post by @gdb
9. X post by @gdb
10. X post by @sama
11. First impressions of GPT-5.5 from Will Koh

12. First impressions of GPT-5.5 from Aaron Friel
13. X post by @gdb
14. X post by @OpenAIDevs
15. X post by @sama
16. OpenAI's New GPT-5.5 Powers Codex on NVIDIA Infrastructure — and NVIDIA Is Already Putting It to Work
17. X post by @deepseek_ai
18. X post by @EMostaque
19. X post by @GoogleDeepMind
20. X post by @GoogleDeepMind
21. X post by @GoogleDeepMind
22. X post by @GoogleDeepMind
23. X post by @GoogleDeepMind
24. X post by @JeffDean
25. X post by @JeffDean