GPT-Live Arrives, Grok 4.5 Lands, and Coding Evals Come Under Scrutiny

AI News Digest

2026-07-09

GPT-Live Arrives, Grok 4.5 Lands, and Coding Evals Come Under Scrutiny

By AI News Digest • July 9, 2026

OpenAI's full-duplex GPT-Live rollout and Grok 4.5's release led the day's product news. Beneath the launches, OpenAI challenged a core coding benchmark, while Anthropic, Modal, and NVIDIA signaled where safety and agent infrastructure are heading.

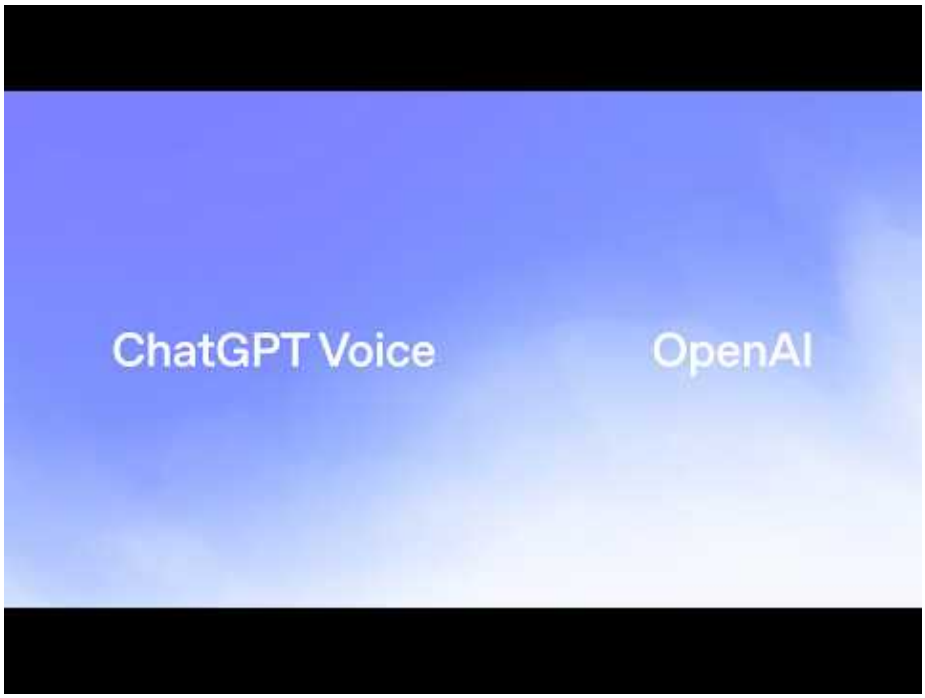
Two launches set the tone

OpenAI rolls out GPT-Live in ChatGPT

OpenAI launched GPT-Live, a new generation of voice models rolling out in ChatGPT across iOS, Android, and web, with API access coming soon [1, 2]. The company said its full-duplex architecture lets the model listen and speak at the same time, handle interruptions more naturally, maintain better time awareness, and perform live translation [3, 4, 5]. For harder tasks, GPT-Live can delegate web search and deeper reasoning to frontier models behind the scenes; the launch video specifically described GPT 5.5 doing that work in parallel while the conversation continues, and OpenAI later said the product was fully rolled out to Go, Plus, and Pro users, with free rollout still in progress [6, 7, 8].

> "i have always preferred typing to talking to an AI, now i think that's going to shift." [9]

Why it matters: This is more than a voice refresh. OpenAI is positioning spoken interaction as a primary interface, not just a voice layer on top of turn-based chat [7, 3].



The next generation of ChatGPT Voice (10:11)

Grok 4.5 arrives with a coding-first pitch

Grok 4.5 was presented as a model for real-world engineering, and by later in the day it was available in Cursor and to all Vercel customers [10, 11, 12]. The model runs on V9, a 1.5 trillion-parameter foundation model, is priced at $2/$6 per 1M input/output tokens, and Musk said its context window will upgrade to 1M next week [13, 14, 15]. Artificial Analysis placed it #4 on its Intelligence Index and GDPval-AA v2, and said it scores 76 on the Coding Agent Index at lower cost and token use than several peers [14].

Why it matters: The launch reinforces a competitive lane centered on coding agents, speed, and cost efficiency rather than a general-purpose consumer assistant pitch [13, 16].

Under the surface, measurement and safety shifted

OpenAI retracts SWE-Bench Pro as a leading frontier coding eval

OpenAI said SWE-Bench Pro is saturated at roughly a 70% noise ceiling; in a later post, it said about 30% of tasks are broken, and it is retracting its earlier recommendation that the field use the benchmark as a leading coding eval [17, 18]. The company said hidden requirements, contradictory instructions, overly strict tests, and incomplete grading criteria distort results, and that its audit

combined model-based investigator agents with reviews from five independent experienced software engineers [19, 20, 21]. OpenAI said stronger coding models now require harder, fairer, and more trustworthy evaluations [22, 23].

Why it matters: Benchmark scores are increasingly central to model launches, so a public downgrade of one of the best-known coding evals should make leaderboard comparisons easier to question [14, 17].

Anthropic backs a modular safety approach

Anthropic said it collaborated with AE Studio on GRAM, a training method that places dual-use capabilities such as virology into removable modules [24, 25]. Anthropic linked to a research note with more detail on "off-switch" dual-use capabilities [25].

Why it matters: The work points toward a safety strategy based on isolating and selectively disabling risky capabilities, rather than treating all behavior as one inseparable model package [24, 25].

The agent stack kept getting more concrete

Modal raises $355M and shifts from DX to AX

Modal disclosed a $355 million Series C and described itself as building an "agent cloud," with internal focus shifting from developer experience to agent experience [26]. The company highlighted elastic inference for custom models, sandboxes that can scale to 100,000 instances for rollouts, and production features such as private networking, multi-node training, and observability [26]. It said its capacity pool spans 17 cloud providers and that it wants to be a specialized sandbox provider rather than a managed-agent layer [26].

Why it matters: This is a sizable funding signal that control over runtime, scaling, and observability is becoming a core part of the agent platform battle [26].

NVIDIA and LangChain package an open enterprise agent stack

NVIDIA said LangChain tuned its Deep Agents harness for Nemotron 3 Ultra, yielding the highest accuracy among open models, more completed tasks at higher throughput, and 10x lower inference cost per run than leading closed models [27]. NVIDIA also said Nemotron 3 Ultra reached business-task parity with the highest-scoring closed models without retraining, and that NemoClaw packages Deep Agents, Nemotron 3 Ultra, and the OpenShell secure runtime into an enterprise blueprint [27]. Jensen Huang framed the broader goal as enabling enterprises to build domain-specific, proprietary AI systems they can control and improve over time [28].

Why it matters: The emphasis is moving beyond frontier-model access alone toward owning the harness, runtime, and domain context around specialized agents [28].

Sources

1. X post by @OpenAI
2. X post by @OpenAI
3. X post by @OpenAI
4. X post by @OpenAI
5. This is the new ChatGPT Voice, powered by GPT-Live
6. X post by @OpenAI
7. The next generation of ChatGPT Voice
8. X post by @OpenAI
9. X post by @sama
10. X post by @SpaceXAI
11. X post by @elonmusk
12. X post by @rauchg
13. X post by @tetsuoai
14. X post by @ArtificialAnlys
15. X post by @elonmusk
16. X post by @elonmusk
17. X post by @OpenAI
18. X post by @OpenAI
19. X post by @OpenAI
20. X post by @OpenAI
21. X post by @OpenAI
22. X post by @OpenAI
23. X post by @OpenAI
24. X post by @AEStudioLA
25. X post by @AnthropicAI
26. Why AI Infrastructure must evolve for Agent Experience — Akshat Bubna, Modal CTO
27. NVIDIA Nemotron Achieves Benchmark-Leading Performance With LangChain Deep Agents Harness
28. Jensen Huang: Why companies need open agent systems