GPT-5.5 Arrives as DeepSeek Opens V4 and Agents Get More Practical

AI High Signal Digest

2026-04-24

GPT-5.5 Arrives as DeepSeek Opens V4 and Agents Get More Practical

By AI High Signal Digest • April 24, 2026

OpenAI shipped GPT-5.5 with an efficiency-focused story, while DeepSeek countered with open-source V4 models built for 1M-token context and aggressive pricing. The brief also covers new agent-infrastructure research, orchestration tools, and a U.S. warning on industrial-scale AI distillation.

Top Stories

Why it matters: The day's biggest signal was a two-front competition: OpenAI pushed a more efficient frontier model into products, while DeepSeek answered with a large open-weight release built around million-token context.

- **OpenAI launched GPT-5.5 in ChatGPT and Codex, with API access coming soon.** OpenAI says the model is built for real work and agents, matches GPT-5.4 per-token latency, and uses significantly fewer tokens on Codex tasks [1, 2]. API pricing is **$5 / 1M input** and **$30 / 1M output** with a **1M context window** [3]. Official results highlighted **82.7%** on Terminal-Bench 2.0 and **81.8%** on CyberGym [4, 5]. Artificial Analysis says GPT-5.5 now leads its Intelligence Index by 3 points, though its AA-Omniscience hallucination rate remains high at **86%** [6].
- **DeepSeek open-sourced V4 Pro and Flash.** The new family ships with **V4-Pro** at **1.6T total / 49B active** and **V4-Flash** at **284B / 13B active**, both with **1M context** and live API/web availability [7]. DeepSeek says V4 uses token-wise compression plus DeepSeek Sparse Attention to cut long-context compute and memory costs [8]. Pricing is aggressive: **$1.74/$3.48** for Pro and **$0.14/$0.28** for Flash per 1M input/output tokens [9]. vLLM says the V4 design reduces per-layer KV state at 1M context by about **8.7x**, and Artificial Analysis already ranks

V4 Pro as the top open-weights model on GDPval-AA [10, 9].

Research & Innovation

Why it matters: The most useful research today focused on making training and agent systems more scalable, while clarifying where multi-agent hype breaks down.

- **Google DeepMind's Decoupled DiLoCo** enables training across multiple data centers without stalling on failures; Google says it trained a **12B Gemma** model across **four U.S. regions** and mixed **TPU v6e/v5p** hardware without performance loss [11, 12, 13, 14].
- **Neural Garbage Collection (NGC)** trains models to manage their own KV cache with RL/GRPO, aiming to stabilize memory for long-horizon reasoning, agents, and tool use [15, 16].
- A new paper on **diversity collapse** found that multi-agent LLM systems can converge toward near-identical outputs over time because shared context and mutual feedback homogenize the group [17].

Products & Launches

Why it matters: New launches are less about one-off demos and more about dependable autonomy, memory, and orchestration.

- **Codex** added broader browser/computer support plus **auto-review**, letting the agent keep moving through tests, builds, files, and UI tasks while a separate checker reviews higher-risk actions [18, 19, 20, 21].
- **Sakana Fugu** entered beta as a multi-agent orchestration system; Sakana says it has hit SOTA on **SWE-Pro, GPQA-D, and ALE-Bench** and ships as an **OpenAI-compatible API** with **Mini** and **Ultra** modes [22].
- **Claude Managed Agents Memory** is now in public beta, giving Anthropic-managed agents a memory layer that learns from prior sessions [23].

Industry Moves

Why it matters: Competition is shifting toward distribution, enterprise rollout, and on-device deployment.

- OpenAI and **NVIDIA** piloted a company-wide **Codex** rollout, and OpenAI says it is now offering whole-company deployments to other enterprises [24, 25].
- **Liquid AI** signed a multi-year **Mercedes-Benz** partnership to bring embedded speech, language understanding, and reasoning into future MBUX systems [26, 27].
- **Glif V2** launched alongside a **$17.5M seed** led by **a16z** and **USV**, positioning itself as a creative super agent for ads, films, voiceovers, and more [28].

Policy & Regulation

Why it matters: The clearest government action today was around model theft and distillation.

- A U.S. memo said foreign entities, primarily in China, are running **industrial-scale distillation campaigns** against American AI and said the government will act to protect domestic innovation [29].

Quick Takes

Why it matters: These are smaller updates, but each points to where competition is moving next.*

- **Kimi K2.6** is now the **#1 open model** in both **Vision Arena** and **Document Arena** [30, 31].
- **Qwen3.6-27B** can run locally on **18GB RAM** and beats the much larger **Qwen3.5-397B-A17B** on major coding benchmarks [32].
- **Kling 3.0** is live with **native 4K video** generation, without upscaling [33].
- Anthropic says recent **Claude Code** quality issues were traced to three problems, fixed in **v2.1.116+**, with usage limits reset for subscribers [34].

Sources

1. X post by @OpenAI
2. X post by @OpenAI
3. X post by @sama
4. X post by @OpenAIDevs
5. X post by @OpenAIDevs
6. X post by @ArtificialAnlys
7. X post by @deepseek_ai
8. X post by @deepseek_ai
9. X post by @ArtificialAnlys
10. X post by @vllm_project
11. X post by @GoogleDeepMind
12. X post by @GoogleDeepMind
13. X post by @GoogleDeepMind
14. X post by @GoogleDeepMind
15. X post by @cwolferesearch
16. X post by @cwolferesearch
17. X post by @dair_ai
18. X post by @OpenAIDevs
19. X post by @OpenAIDevs
20. X post by @OpenAIDevs
21. X post by @gdb

22. X post by @SakanaAILabs
23. X post by @claudeai
24. X post by @sama
25. X post by @gdb
26. X post by @liquidai
27. X post by @maximelabonne
28. X post by @fabianstelzer
29. X post by @mkratsios47
30. X post by @arena
31. X post by @Kimi_Moonshot
32. X post by @UnslothAI
33. X post by @openart_ai
34. X post by @ClaudeDevs