OpenAI Moves on Ona, Gemini Tops Video, and the AI Price War Intensifies

AI High Signal Digest

2026-06-12

OpenAI Moves on Ona, Gemini Tops Video, and the AI Price War Intensifies

By AI High Signal Digest • June 12, 2026

OpenAI moved to strengthen Codex with an acquisition, Gemini took a clear lead in video generation benchmarks, and the business fight between Anthropic and OpenAI is increasingly about pricing and margins. The brief also covers a sobering new agent benchmark, automated AI research, and key product and infrastructure launches.

Top Stories

Why it matters: today's clearest signals were about agent infrastructure, multimodal model performance, and the economics of the frontier-model race.

- **OpenAI moved to strengthen Codex's agent stack** by reaching an agreement to acquire Ona. OpenAI said Ona's secure cloud execution technology will help Codex handle longer-running work even when laptops are closed and help organizations deploy agents securely in production; Ona will join the Codex team after closing. [1]
- **Gemini Omni Flash jumped to the top of video generation benchmarks.** It now ranks #1 in Video Arena for both text-to-video and image-to-video, with Arena reporting a +158 point gain over Veo 3.1 (1080p) in text-to-video, a +61 point lead over Seedance 2.0, a +77 point gain in image-to-video, and an 82% head-to-head win rate. DeepMind has framed Omni as its first step toward a model that can "create anything from anything," starting with video. [2, 3, 4, 5]
- **Competition is shifting from benchmarks to margins.** Anthropic is reportedly on track for its first profitable quarter, with revenue more than doubling to about $10.9B, while OpenAI is reportedly weighing further token price cuts to keep enterprises from moving to Claude. Separate

estimates put $200 plans at roughly $8,000 of Claude Max 20x usage and $14,000 of ChatGPT Pro 20x usage, highlighting how aggressive current pricing already is. [6, 7]

Research & Innovation

Why it matters: the strongest research updates sharpened the picture on agent limits, automated research, and training-data control.

- **Agents' Last Exam (ALE)** introduced a rolling benchmark built from more than 1,500 expert-sourced tasks across 55 occupations. The results were mixed: today's frontier agents can solve a meaningful fraction of professional tasks, but every frontier system tested scored **0%** on the hardest tier requiring sustained reasoning, deep expertise, and long-horizon execution. Separate commentary said GPT-5.5 led the eval even when measured by tokens, cost, or wall-clock time. [8, 9]

  > "The age of useful agents is here. The age of truly job-ready agents is not." [8]

- **Recursive** unveiled an early automated AI research system, "Eureka Machine" v0.1, that it says reached state-of-the-art results on NanoGPT speedrun, NanoChat, and NVIDIA's Sol-ExecBench. The company says the code and ideas behind those results were generated by the AI system itself and is open-sourcing the discoveries for scrutiny. [10]
- **Goodfire introduced "predictive data debugging,"** a method for estimating what DPO training data will amplify or suppress before training, reporting R^2=0.9 against what models later learn. In examined datasets, it surfaced weaker safety guardrails, hallucinated links on sensitive topics, and localized sycophancy. [11, 12, 13, 14]

Products & Launches

Why it matters: new releases kept pushing agents, open coding models, and media systems closer to practical deployment.

- **Perplexity integrated Deep Research directly into Computer.** The feature is built on a "Search as Code" architecture in which the model writes code to assemble search, runs thousands of retrieval steps in parallel, connects to long-running sandboxes and tools, and is now available to Pro and Max subscribers. Perplexity says it outperforms its legacy Deep Research on every benchmark. [15, 16]
- **Cohere released North Mini Code 1.0,** its first open-source coding model. It is a 30B-parameter MoE with 3B active parameters running at about 66 tok/s in BF16, with day-zero MLX support and local deployment paths through GGUF quants, llama.cpp, Ollama, and vLLM. [17, 18, 19]
- **Ideogram 4.0 became Ideogram's first open-weights release.** It debuted at #8 on the Open Weights Text-to-Image leaderboard and #31

overall, with 2K x 2K outputs, multilingual text rendering, bounding-box layout control, transparent backgrounds, and structured JSON prompts. [20]

Industry Moves

Why it matters: capital and infrastructure decisions are increasingly defining who can build, train, and deploy the next generation of systems.

- **Jeff Bezos raised $12B for Prometheus** at a $41B valuation. The company's pitch is an "artificial general engineer" that compresses the design-to-build loop by 10x or more, alongside a reported $100B vehicle to acquire industrial companies and the manufacturing data they generate. [21]
- **xAI is building a 500 MW data center in Saudi Arabia** with HUMAIN and NVIDIA, which would make it xAI's largest facility outside the U.S.; for comparison, Colossus-1 in Memphis is around 300 MW. [22]
- **Google DeepMind launched a $10M research fund** with partners including Schmidt Sciences, Cooperative AI, and ARIA to study how AI systems behave as a group. [23]

Policy & Regulation

Why it matters: governments are starting to define how AI connects to national infrastructure, not just consumer software.

- **China's MIIT issued an AI+ICT implementation plan for 2026-2028** that ties together the "East Data West Compute" project, autonomous networks, agentic AI, embodied intelligence, and the national compute-network strategy. [24]

Quick Takes

Why it matters: these smaller updates still add signal on deployment speed, tool quality, and inference economics.

- Trajectory Labs says it post-trained NVIDIA Nemotron 3 Ultra on Harvey Legal Agent Bench in under 24 hours, putting an open model into the same performance band as leading closed legal models at lower cost. [25]
- Cursor made Auto-review the default for new users; its classifier subagent reviews actions in context and reportedly hits 97% accuracy. [26]
- Baseten and Inception launched Mercury 2 in production, citing 1,000+ tok/s on standard NVIDIA GPUs and early reports of 82% lower latency and 90% cost savings. [27]
- OpenAI added developer mode for Codex browser use in Chrome and the in-app browser, with Chrome DevTools Protocol support for profiling JavaScript, console output, network traffic, and page state. [28]

Sources

1. X post by @OpenAINewsroom
2. X post by @arena
3. X post by @arena
4. X post by @arena
5. X post by @GoogleDeepMind
6. X post by @kimmonismus
7. X post by @kimmonismus
8. X post by @dawnsongtweets
9. X post by @polynoamial
10. X post by @RichardSocher
11. X post by @GoodfireAI
12. X post by @GoodfireAI
13. X post by @GoodfireAI
14. X post by @GoodfireAI
15. X post by @perplexity_ai
16. X post by @perplexity_ai
17. X post by @Prince_Canuma
18. X post by @cohere
19. X post by @cohere
20. X post by @ArtificialAnlys
21. X post by @kimmonismus
22. X post by @kimmonismus
23. X post by @GoogleDeepMind
24. X post by @pstAsiatech
25. X post by @trajectorylabs
26. X post by @cursor_ai
27. X post by @baseten
28. X post by @OpenAIDevs