OpenAI's Jalapeño Chip Leads a Day of Agent and Infrastructure Shifts

AI High Signal Digest

2026-06-25

OpenAI's Jalapeño Chip Leads a Day of Agent and Infrastructure Shifts

By AI High Signal Digest • June 25, 2026

OpenAI's custom Jalapeño chip led the day, alongside Gemini's new computer-use mode and a $200M launch for MirendilAI. Elsewhere, GLM-5.2 advanced open-model benchmarks, while the first legal challenge to US AI export controls moved policy risk into court.

Top Stories

Why it matters: today's clearest signals were about who controls AI compute, how far agents can act on software surfaces, and where new frontier R&D money is going.

- **OpenAI unveiled Jalapeño, its first custom inference chip.** Built with Broadcom for ChatGPT, Codex, the API, and future agentic products, it extends OpenAI's stack into infrastructure [1]. OpenAI said the program went from initial design to tape-out in nine months, used ChatGPT in the engineering process, is already running GPT-5.3-Codex-Spark in the lab, and targets substantially better performance per watt, with deployment planned for end-2026 [2, 3]. The stated goal is lower dependence on external GPUs and tighter control over compute economics [3].
- **Google launched computer use in Gemini 3.5 Flash.** The feature lets an agent take a screen and a goal, then determine actions across browser, mobile, and desktop environments [4]. Shared examples show it auditing docs pages by navigating, running code snippets, taking screenshots, and returning a report, with safeguards including user confirmation and auto-stop on prompt injection [4].
- **MirendilAI launched with a $200M seed round.** The startup says

it will build self-accelerating AI R&D systems, with a 20-person founding team from Anthropic, xAI, Google DeepMind, and OpenAI; the round was led by a16z and Kleiner Perkins, with a major NVIDIA investment [5]. Its pitch is to democratize frontier AI capabilities for broader scientific use rather than concentrate them in a few labs [5].

Research & Innovation

Why it matters: the most notable technical updates focused on open-model progress, learning limits, and better ways to compose agents.

- **GLM-5.2 posted the strongest ARC-AGI-2 result yet for an open-source model.** Verified scores were 22.8% on ARC-AGI-2 at $0.25 and 77.0% on ARC-AGI-1 at $0.19, with ARC Prize saying performance is comparable to GPT-5.4 and GPT-5.5 at low reasoning effort [6]. François Chollet called it the strongest ARC-AGI-2 performance to date by an open-source model [7].
- **Zyphra argued that continual learning hits a deeper failure mode than forgetting.** Its new work identifies plasticity loss as models losing the ability to learn new data, shows it across 5M-314M parameter GPT-style models, and reports the same decline even in stationary pretraining [8, 9, 10]. The team fit a scaling law for the onset, T P^0.83, suggesting scale delays the problem but does not remove it [11].
- **AI21 topped DeepResearch Bench II by merging weak agents instead of building a new one.** It combined seven agents ranked 7-13 into a single report pipeline and reported a new #1 score of 64.38 [12, 13].

Products & Launches

Why it matters: product updates kept pushing models from chat into domain tools and team workflows.

- **OpenAI updated GPT-5.5 Instant.** The new version is described as better at understanding user intent, adapting responses, handling complex constraints, and improving shopping and local recommendations; rollout started today for paid users and tomorrow for free users [14].
- **Perplexity launched Computer for Counsel.** The product connects legal research databases, document tools, and matter-management systems so lawyers can pull citable sources from tools including MidpageAI, LegalZoom, Docusign, and NetDocuments [15]. It is available to Pro and Max subscribers [15].
- **Notion and Cursor pushed agents deeper into team workflows.** Notion introduced External Agents with Claude and Cursor so teams can assign work from shared boards and @-mention agents like teammates [16].

Cursor said the integration runs on its SDK so cloud agents can take tasks from Notion and open PRs using the same runtime as Cursor itself [17].

Industry Moves

Why it matters: infrastructure control and talent concentration remain central competitive levers.

- **Qualcomm agreed to acquire Modular.** Both sides said the deal is meant to unify accelerated compute with an open platform spanning edge to cloud and hardware from CPUs and GPUs to NPUs and custom ASICs [18, 19, 20].
- **Anthropic is pulling more talent from Google DeepMind.** Bloomberg reported that Jonas Adler and Alexander Pritzel, both viewed internally as key Gemini contributors, are leaving Google for Anthropic [21].

Policy & Regulation

Why it matters: AI governance is shifting from abstract debate toward concrete fights over access and supply-chain alignment.

- **The first legal challenge to the Trump administration's AI export controls has arrived.** Legion is suing over the forced shutdown of Anthropic's Fable 5 and Mythos 5 for foreign nationals, arguing export-control laws do not cover access to hosted AI models or text outputs and that no national emergency was declared [22, 23]. The case turns on whether hosted frontier-model access can be treated as export-controlled technology when users only receive text outputs [22].
- **Europe joined a US-led AI supply-chain pact.** The EU, Germany, the Netherlands, and Greece joined Pax Silica, covering chips, critical minerals, energy, and compute; Jacob Helberg explicitly positioned it against digital sovereignty built around duplicative national tech stacks [24].

Quick Takes

Why it matters: these smaller updates still show where deployment patterns are heading.

- Anthropic's new **agent identity** model gives Claude its own credentials in shared channels, while DMs run on the user's connectors, with one auditable identity for admins [25, 26, 27, 28].
- **Google AI Studio** says more than **1 million Android apps** have been created since native Android app building launched in May [29].
- **Wan-2.7 I2V** entered Video Arena at **#5**, ahead of Grok Imagine Video and every Google Veo-3.1 variant [30].

- **Kog** open-sourced the 2B **Laneformer** model used to demonstrate **3,000+ tokens per second** [31].

Sources

1. X post by @OpenAI
2. X post by @kimmonismus
3. X post by @kimmonismus
4. X post by @_philschmid
5. X post by @bneyshabur
6. X post by @arcprize
7. X post by @fchollet
8. X post by @ZyphraAI
9. X post by @ZyphraAI
10. X post by @ZyphraAI
11. X post by @ZyphraAI
12. X post by @AI21Labs
13. X post by @AI21Labs
14. X post by @OpenAI
15. X post by @perplexity_ai
16. X post by @NotionHQ
17. X post by @cursor_ai
18. X post by @Qualcomm
19. X post by @clattner_llvm
20. X post by @clattner_llvm
21. X post by @EdLudlow
22. X post by @kimmonismus
23. X post by @SophiaCai99
24. X post by @kimmonismus
25. X post by @ClaudeDevs
26. X post by @ClaudeDevs
27. X post by @ClaudeDevs
28. X post by @ClaudeDevs
29. X post by @GoogleAIStudio
30. X post by @arena
31. X post by @ClementDelangue