Anthropic Locks In SpaceX Compute as OpenAI Opens Its Training Network Stack

AI High Signal Digest

2026-05-07

Anthropic Locks In SpaceX Compute as OpenAI Opens Its Training Network Stack

By AI High Signal Digest • May 7, 2026

Anthropic turned a major SpaceX compute partnership into higher Claude limits, while OpenAI opened up the networking protocol behind its largest training clusters. The brief also covers Zyphra's compact reasoning model, DeepSeek's fundraising talks, new agent tools, and a possible U.S. pre-release model review regime.

Top Stories

Why it matters: The clearest signal today is that AI competition is being shaped as much by infrastructure access as by model quality.

- **Anthropic's SpaceX deal is already changing Claude capacity.** Anthropic said its partnership with SpaceX will substantially increase compute capacity, including all compute capacity at the Colossus 1 data center and more than 300 megawatts deployable within a month [1, 2]. The company tied that capacity directly to higher usage limits for Claude Code and the Claude API, and said Claude inference on Colossus will begin ramping in the next few days [1, 3]. Separately, Elon Musk said xAI will be dissolved as a separate company into **SpaceXAI**, while xAI said SpaceXAI and Anthropic have expressed interest in developing multiple gigawatts of orbital AI compute [4, 5].
- **OpenAI released part of the networking stack behind frontier training.** OpenAI, together with AMD, Broadcom, Intel, Microsoft, and NVIDIA, launched Multipath Reliable Connection (MRC), an open protocol meant to make large AI training clusters faster, more reliable, and less wasteful of GPU time [6]. OpenAI says MRC is already deployed on its largest frontier-model supercomputers, including OCI Abilene and Mi-

crosoft Fairwater, and is now available through Open Compute for others to build on [7].

Research & Innovation

Why it matters: The most useful research updates today were about model efficiency, retrieval limits, and speeding up reinforcement learning.

- **Zyphra's ZAYA1-8B is a notable open-model release.** Zyphra released ZAYA1-8B, a reasoning MoE trained on AMD and optimized for high intelligence density [8]. The company says it uses fewer than 1B active parameters yet beats open-weight models many times its size on math and reasoning, approaching DeepSeek-V3.2 and GPT-5-High with test-time compute [8].
- **OBLIQ-Bench goes after a real retrieval bottleneck.** Researchers built the benchmark after finding little headroom left in many hard IR benchmarks even with oracle reranking by frontier LLMs [9]. Its core idea is to test cases where reasoning models can recognize subtle relevance once shown a document, but scalable retrieval systems still fail to surface that document from the corpus [10].
- **NVIDIA showed speculative decoding can speed up RL without changing model behavior.** A new result reports up to 2.5x faster end-to-end reinforcement learning at 235B scale, while keeping the final sampled sequence consistent with the original large model's distribution [11]. The team also reports roughly 1.8x faster rollout throughput at 8B scale in a full NeMo-RL + vLLM pipeline [11].

Products & Launches

Why it matters: Product releases focused on better agent inputs: better data, better grounding, and better memory.

- **Perplexity added licensed finance data to its Agent API.** Finance Search gives developers one-call access to licensed financial datasets, live market data, and cited web sources for tasks like valuation lookups, earnings recaps, and market monitoring [12, 13]. Perplexity says it achieved the highest accuracy for live financial data and the lowest cost per correct answer on FinSearchComp T1 [14].
- **Google is making AI Search more link-rich.** Updates to AI Mode and AI Overviews add more article suggestions, inline links, subscription-source highlighting, desktop hover previews, and previews of discussions and social sources with creator context [15].
- **Claude's new Dreaming feature pushes agents toward longer-term memory.** Anthropic says Dreaming reviews past agent sessions, extracts patterns, and curates memories so agents can learn over time [16].

Industry Moves

Why it matters: Capital, defense demand, and strategic research partnerships are still concentrating around a small number of AI players.

- **Scale AI deepened its Pentagon footprint.** The company won a $500 million DoD contract through the Chief Digital and AI Office to help sift data and assist decision-making, following a $100 million deal in 2025 [17].
- **DeepSeek is reportedly nearing a $45 billion raise.** Multiple reports say the company is in talks for its first fundraising round at roughly that valuation, with China's largest state-backed semiconductor fund involved and investors betting on commercialization of DeepSeek's coding strength despite an undeveloped business model [18, 19].
- **DeepMind is turning EVE Online into an AI research sandbox.** Google DeepMind said EVE's player-driven universe is a strong environment for testing memory, continual learning, and long-term planning, and Bloomberg separately reported Google took a multi-million-dollar stake in the game's developer [20, 21].

Policy & Regulation

Why it matters: There is one policy signal that could matter a lot if it hardens into an actual release gate.

- **The White House is reportedly considering an FDA-like model vetting process.** Reporting says the administration is weighing an executive order to review new AI models for safety before release [22]. No finalized action was cited in the notes, so this remains a proposal rather than a rule.

Quick Takes

Why it matters: These smaller updates still sharpen the competitive picture.

- **Harvey's LAB** is positioned as a 1,200-task legal-agent benchmark spanning 24 practice areas, with Artificial Analysis partnering to track results [23].
- **Google Translate Live translate** now offers real-time translations in 70+ languages through any headphones [24].
- **OpenAI Codex subagents** can split work across specialized agents and recombine results for larger codebases and PR reviews [25].
- **Gemini API File Search** now supports multimodal retrieval for PDFs and images with a single call [26].

Sources

1. X post by @claudeai

2. X post by @claudeai
3. X post by @nottombrown
4. X post by @elonmusk
5. X post by @xai
6. X post by @OpenAI
7. X post by @OpenAI
8. X post by @ZyphraAI
9. X post by @dianetc_
10. X post by @lateinteraction
11. X post by @TheTuringPost
12. X post by @perplexity_ai
13. X post by @perplexity_ai
14. X post by @perplexity_ai
15. X post by @Google
16. X post by @claudeai
17. X post by @Techmeme
18. X post by @jukan05
19. X post by @zijing_wu
20. X post by @GoogleDeepMind
21. X post by @cecianasta
22. X post by @Polymarket
23. X post by @ArtificialAnlys
24. X post by @Google
25. X post by @reach_vb
26. X post by @_philschmid