Anthropic's Colossus Deal and OpenAI's MRC Release Put Compute Center Stage

AI News Digest

2026-05-07

Anthropic's Colossus Deal and OpenAI's MRC Release Put Compute Center Stage

By AI News Digest • May 7, 2026

Compute dominated the day: Anthropic moved to absorb all of Colossus 1 while OpenAI opened a networking protocol already running on its biggest training clusters. Also in view: Moonshot AI's large new round, DeepSeek 4's long-context efficiency push, and a pair of notable robotics launches.

What stood out

Compute was easily the day's center of gravity. One frontier lab locked up an entire training site, while another published the networking layer it uses to keep giant clusters stable at scale.

Anthropic takes all capacity at Colossus 1

Anthropic said its agreement with SpaceX means it will use all compute capacity at the Colossus 1 data center, adding more than 300 megawatts of capacity that can be deployed within the month [1]. NVIDIA described the partnership as powered by 220,000+ GPUs inside Colossus 1 [2]. Musk said he approved leasing Colossus 1 to Anthropic after spending time with senior Anthropic staff and after SpaceXAI had moved training to Colossus 2; in a separate post, he said xAI will be dissolved as a separate company and folded into "SpaceXAI" [3, 4].

Why it matters: This is an unusually explicit sign that frontier-model competition is being shaped not just by model releases, but by who controls and reallocates large-scale compute capacity.

OpenAI turns cluster networking into a product for the wider industry

OpenAI, alongside AMD, Broadcom, Intel, Microsoft, and NVIDIA, released Multipath Reliable Connection (MRC), an open networking protocol meant to make large AI training clusters faster and more reliable with less wasted GPU time [5]. OpenAI said MRC is already deployed across its largest supercomputers, including Oracle Cloud Infrastructure's site in Abilene and Microsoft's Fairwater systems, and is now available through the Open Compute Project for wider industry use [6]. Supporting technical material from NVIDIA and OpenAI described MRC as load-balancing traffic across multiple paths, bypassing failures in hardware at very high speed, and reducing GPU idle time during congestion or link failures [7, 8].

Why it matters: The broader point, echoed by Greg Brockman, is that AI bottlenecks are no longer just about buying more GPUs; they are increasingly about making networking, storage, scheduling, and reliability work together at frontier scale [9].

Moonshot AI adds another large funding signal from China

Moonshot AI's Kimi is closing a $2 billion round at a $20 billion+ post-money valuation, led by Meituan Dragonball with China Mobile and CPE also participating [10]. The cited report says the company's ARR rose from $100 million in early March to more than $200 million by April, driven by subscriptions and API usage, and that total fundraising now exceeds $3.9 billion [10]. The same report called Kimi the most-funded Chinese AI startup so far [10].

Why it matters: This was one of the clearest capital signals of the day: major money is still flowing to model companies that can pair rapid revenue growth with large strategic investors.

DeepSeek 4 keeps pushing the open-model race toward long context and efficiency

A Two Minute Papers review of DeepSeek 4 highlighted a 58-page paper describing an open-weight model with a 1 million-token context window, enough for roughly 1,500 pages of dense documentation [11]. The review said the model uses several forms of KV-cache compression, with the Pro version requiring about 3x less compute than its predecessor during generation and the Flash version about 10x less [11]. It also cited reported long-context recall results above Gemini 3.1 Pro, while noting that the system is text-only and degrades near its context limits [11].

Why it matters: The signal here is not only benchmark performance. Open models are competing on how efficiently they can use very large context windows, not just on raw scale.

Robotics news split between easier building and deeper embodiment

Hugging Face launched an "agentic robotics app store" for Reachy Mini, saying 300+ apps have shipped and 10,000 robots are already in the wild; it framed the goal as making robotics app development possible in hours rather than weeks, including for non-coders [12]. Separately, gs_ai_ introduced GENE-26.5, a "robotic brain" built around a robotics-native foundation model, a 1:1 human-like hand, a noninvasive glove for motion, force, and touch data, and a simulator, with one model handling language, vision, proprioception, tactile input, and action [13].

Why it matters: Taken together, these launches point in two complementary directions for robotics: lowering the barrier to shipping robot applications and building richer full-stack systems for embodied AI.

Sources

1. X post by @claudeai
2. X post by @nvidia
3. X post by @elonmusk
4. X post by @elonmusk
5. X post by @OpenAI
6. X post by @OpenAI
7. NVIDIA Spectrum-X — the Open, AI-Native Ethernet Fabric — Sets the Standard for Gigascale AI, Now With MRC
8. Why AI needs a new kind of supercomputer network — the OpenAI Podcast Ep. 18
9. X post by @udayruddarraju
10. X post by @ChengManqi
11. DeepSeek V4 AI Beats Billion Dollar Systems…For Free
12. X post by @ClementDelangue
13. X post by @gs_ai_