Anthropic's Reported Surge, NVIDIA's 4-Bit Breakthrough, and ChatGPT's Finance Push

AI High Signal Digest

2026-05-16

Anthropic's Reported Surge, NVIDIA's 4-Bit Breakthrough, and ChatGPT's Finance Push

By AI High Signal Digest • May 16, 2026

Reports this week pointed to a sharp jump in Anthropic's scale, NVIDIA showed frontier-style training in 4-bit precision, and OpenAI pushed ChatGPT closer to a personal agent with connected finance data. Also in view: new reasoning research, developer tools, and major funding and infrastructure signals.

Top Stories

Why it matters: the biggest signals today were commercial scale, cheaper frontier training, and assistants moving closer to acting on personal context.

- **Anthropic's reported economics jumped again.** A Financial Times-linked post pegged Anthropic at a **$900B valuation**, up from **$350B** in February, and said ARR rose from **$9B** at the end of 2025 to **$45B** by the end of May. Separate interview notes this week also said Anthropic has **9 of the Fortune 10** as customers and **$100B** in combined compute commitments. Together, those figures point to how quickly enterprise AI spending is concentrating around a few frontier labs. [1, 2, 3]
- **NVIDIA pushed 4-bit training from an efficiency trick toward a frontier-scale method.** NVIDIA said it trained a **12B** parameter LLM in **NVFP4** on **10T tokens** with near-zero intelligence loss, matching 8-bit baselines on MMLU, GSM8K, and coding benchmarks. The company also said NVFP4 delivers **2x-3x faster arithmetic**, **50% lower memory use**, and has already been used to pretrain **120B** and roughly **500B** Nemotron models. [4, 5]
- **ChatGPT moved deeper into personal data.** OpenAI launched a personal finance experience for U.S. Pro users that lets them securely

connect financial accounts, view spending, and ask GPT-5.5 questions grounded in transaction data. A follow-up post said the feature uses Plaid, cannot move money or see full account numbers, and is part of the broader push toward ChatGPT as a personal agent for home and work. [6, 7, 8]

Research & Innovation

Why it matters: today's most interesting research updates were about stronger reasoning, better training data, and model reliability.

- **A new reasoning model reached Olympiad-level results.** A **30B-A3B** model was released with gold-medal-level performance on **IPhO** and on **IMO/USAMO** evaluations through test-time self-verification and refinement, alongside what its authors called a simple unified scaling recipe for proof search. [9]
- **FrontierSmith targets the open-ended coding data bottleneck.** The system mutates closed-ended coding tasks into runnable optimization environments for long-horizon agents, and its authors said FrontierSmith-trained models outperformed models trained on human-curated open-ended data on **FrontierCS** and **ALE-bench**. [10]
- **A new fine-tuning result exposed a safety failure mode.** Researchers found that models fine-tuned on documents discussing implausible claims - even when those documents explicitly say the claims are false - can end up believing the claims anyway, raising doubts about how robust some current control methods are. [11, 12]

Products & Launches

Why it matters: new launches were less about flashy chat and more about making agents useful inside real workflows.

- **Cohere launched Compass** for search and retrieval over unstructured data, including handwritten or typed scans and other difficult documents, using a visual parsing model plus an advanced embedding stack. [13]
- **Notion expanded its developer platform for agents.** New additions include agent tools, webhook triggers, an External Agents API, and a Notion Agents SDK, with Notion saying the long-term aim is for users' agents to build workflows for them. [14]
- **VS Code added AI-generated risk badges for terminal commands.** Commands are now labeled as safe, caution, or review carefully before execution, with an experimental setting to enable the feature. [15]

Industry Moves

Why it matters: capital, revenue, and infrastructure scale are now moving almost as fast as the models themselves.

- **Cognition's Devin is showing unusually fast business traction.** Posts this week said Devin reached a **$445M** revenue run rate in its first 18 months, with usage doubling every eight weeks, customers including the **US Army, Goldman Sachs, and Mercedes-Benz**, and a new raise at around a **$25B** valuation. Cognition also said AngelList completed a troubled **14,000-dashboard** migration **5.2x faster** than projected using Devin. [16, 17, 18]
- **Recursive_SI launched with a $650M raise.** The company said more than a third of its team is based in the UK and described its work as contributing to UKSovereignAI goals with UK government support. [19, 20]
- **The AI buildout is becoming a capital-markets story.** One analysis this week said hyperscaler capex is set to cross **$600B** this year, while Big Tech is spending roughly **$400B/year** on AI infrastructure against about **$100B** in AI revenue, highlighting the financing strain behind the current buildout. [21]

Quick Takes

Why it matters: these smaller updates still help map where the ecosystem is heading next.

- xAI said its **Grok V9 1.5T** run is complete and looking strong even before supplemental training with Cursor data. [22]
- Anthropic reset users' **5-hour and weekly Claude limits**. [23]
- **DALL-E 3** will retire from **Bing Image Creator** in the coming weeks; Microsoft says it is building a dedicated replacement. [24]
- **vLLM v0.21.0** added DeepSeek V4 support, speculative decoding that respects reasoning budgets, and NVFP4/MXFP4 quantization, alongside breaking changes including a **C++20** requirement. [25, 26]

Sources

1. X post by @kimmonismus
2. X post by @kimmonismus
3. X post by @patrick_oshag
4. X post by @HowToAI_
5. X post by @ctnzr
6. X post by @ChatGPTapp
7. X post by @kimmonismus

8. X post by @gdb
9. X post by @stingning
10. X post by @MangQiuyang
11. X post by @OwainEvans_UK
12. X post by @RyanPGreenblatt
13. X post by @cohere
14. X post by @NotionHQ
15. X post by @code
16. X post by @colossusmag
17. X post by @cognition
18. X post by @cognition
19. X post by @KanishkaNarayan
20. X post by @_rockt
21. X post by @kimmonismus
22. X post by @elonmusk
23. X post by @ClaudeDevs
24. X post by @JordiRib1
25. X post by @vllm_project
26. X post by @vllm_project