Cognition's $1B+ Raise, Open Protein Models, and the Buildout of Agent Infrastructure

AI News Digest

2026-05-28

Cognition's $1B+ Raise, Open Protein Models, and the Buildout of Agent Infrastructure

By AI News Digest • May 28, 2026

Agents dominated the day, but the signal was less about demos than about economics and deployment: Cognition disclosed major financing and revenue, Trajectory launched a continual-learning bet, and NVIDIA pushed token efficiency to the foreground. On the research side, BioHub released open protein models and Sakana proposed a memory-light training method.

What stood out today

The clearest pattern was AI moving deeper into operating questions: financing at real scale, post-deployment learning, token economics, and system design for production agents [1, 2, 3, 4].

The operating layer got clearer

Cognition pairs a huge round with rare operating metrics

Cognition said it raised over $1B at a $26B valuation led by Lux Capital, General Catalyst, and 8vc. It also said enterprise usage is up more than 10x since the start of the year and run-rate revenue has reached $492M; the company added that cloud agents have gone from niche to mainstream since Devin launched two years ago [1].

Why it matters: The announcement paired financing with concrete usage and revenue metrics, which gives a public read on enterprise demand for coding agents rather than just model hype [1].

Trajectory launches to turn product usage into continual learning

Trajectory launched as a research lab and product company building a platform for continual learning, saying it uses signal from product usage to continuously post-train large-scale agentic models. It launched with $15M in funding, named partners including Clay, Harvey, Decagon, Mercor, and Rogo, said some deployments are already in production, and described a team drawn from DeepMind, OpenAI, Apple, Meta, Amazon AGI, Scale AI, Stripe, and Figma [2].

Why it matters: This is a direct bet that deployed products should keep improving after launch. Nathan Lambert argued continual learning is likely to show up first in knowledge-work products trained on real-world data and RL, while Sarah Guo framed the underlying problem simply: AI is still largely frozen after deployment [5, 6].

NVIDIA sharpens the AI factory pitch around token economics

> "AI factories convert energy into tokens" [3]

NVIDIA is positioning AI factories as infrastructure for always-on reasoning models, agents, and multi-agent systems [3]. It says Blackwell Ultra delivers the lowest cost per token, with GB300 NVL72 systems producing 50x more tokens per megawatt and 35x lower cost per token than Hopper, while Vera Rubin systems are designed for up to 35x higher performance per watt [3].

Why it matters: The story here is full-stack orchestration, not just chips: compute, memory, networking, software, and facility design tuned around token throughput and latency. NVIDIA is tying that pitch to deployment as well, naming Cisco, Dell, HPE, Lenovo, and Supermicro as partners and saying its own enterprise AI factory already uses hundreds of autonomous agents internally [3].

Research releases worth tracking

BioHub open-sources a new protein-model stack

BioHub released the MIT-licensed ESMC protein world model and ESMFold2, alongside an atlas of 6.8B proteins and 1.1B predicted structures. The project says ESMC was trained on 2.8B sequences, added metagenomic data beyond UniRef, and achieved state-of-the-art performance on protein interactions, especially antibodies, with evidence of inference-time scaling across five cancer and immunology targets [7, 8].

Why it matters: The release is a strong argument for scaling large, general protein language models rather than relying only on specialized pipelines. Latent Space highlighted the claim that this approach can beat specialized systems like AlphaFold3 on some hard protein problems, especially where MSAs are weak or unavailable, such as antibodies [7, 8].

Sakana's DiffusionBlocks aims at the training memory wall

Sakana AI Labs introduced DiffusionBlocks, a framework that trains networks one block at a time by interpreting each block as moving representations closer to the target, like a diffusion process. The approach only needs memory for a single block, matched end-to-end training across five architectures, and extends to recurrent-depth transformers by replacing backpropagation through time with a single forward pass during training [9].

Why it matters: The paper is explicitly framed as a response to the resource wall created by end-to-end backprop. If the results hold up, it offers a practical route to reducing training memory without giving up competitive performance [10, 9].

MiniMax's M2 report gives a rare look at production agent training

A new MiniMax M2 technical report argues that full attention still beat hybrid sliding-window variants for production use, and that linear or sparse attention was too fragile when KV-like state is stored in lower precision and when prefix caching matters for coding agents [11, 4]. It also describes an agent-training pipeline built from GitHub pull requests, runnable Docker environments, task-specific rewards, retained reasoning blocks across turns, and wall-clock rewards to discourage slow tool use; the report says self-evolution already handles 30-50% of daily RL iterations and improved internal evaluations by 30% [4].

Why it matters: What stands out is the level of operational detail. The report discusses attention tradeoffs, agent harness design, and RL rewards rather than stopping at top-line benchmarks [4].

Brief notes

- **Genesis World 1.0:** Genesis-Embodied-AI open-sourced a robotics simulation stack built around Genesis World, Quadrants, and Nyx, saying it reaches near-realtime performance, supports contact-rich dexterous manipulation, cuts launch time by 10x, and lowers the sim-to-real gap for zero-shot evaluation [12].
- **Private MCP inside enterprise networks:** OpenAI introduced support for private MCP servers that stay inside a company's network while ChatGPT, Codex, and the Responses API connect through outbound-only HTTPS. The update is aimed at teams that want internal tool access without exposing the MCP server directly [13, 14].
- **Pricing by outcomes, not tokens:** Sierra said it is pricing AI agents on delivered outcomes such as a mortgage completed or a claim settled, rather than token usage. It is an early sign that some vendors want agent pricing tied to business results instead of raw consumption [15, 16, 17].

Sources

1. X post by @cognition
2. X post by @rronak_
3. AI Factories: The New Infrastructure of Intelligence
4. X post by @rasbt
5. X post by @natolambert
6. X post by @saranormous
7. ESMFold2: The Bitter Lesson is Coming for Proteins - Alex Rives, Bio-Hub
8. The Bitter Lesson is Coming for Proteins - Alex Rives, BioHub
9. X post by @SakanaAILabs
10. X post by @hardmaru
11. X post by @RyanLeeMiniMax
12. X post by @gs_ai_
13. X post by @OpenAIDevs
14. X post by @gdb
15. X post by @SierraPlatform
16. X post by @saranormous
17. X post by @saranormous