Cognition Funding, GPT-5.5 Cyber Gains, and the Enterprise Agent Reality Check

AI High Signal Digest

2026-05-28

Cognition Funding, GPT-5.5 Cyber Gains, and the Enterprise Agent Reality Check

By AI High Signal Digest • May 28, 2026

Today's brief centers on Cognition's billion-dollar fundraise, sharp new cybersecurity results for GPT-5.5, and a benchmark showing frontier models are still under 50% on real SRE tasks. It also covers memory-efficient training, protein design, multimodal embeddings, and new enterprise governance controls.

Top Stories

Why it matters: today's clearest signals were where capital is concentrating, where frontier capability is accelerating, and where enterprise agents still fall short.

- **Cognition raised at a new scale.** The company said it raised over $1B at a $26B valuation, with enterprise usage up more than 10x since the start of the year and run-rate revenue at $492M. It also said Devin launched two years ago as the first AI software engineer, and that cloud agents have gone from niche to mainstream [1]. The combination of financing, usage growth, and revenue scale makes this a major commercial signal for coding agents.
- **GPT-5.5 made a large jump on offensive cyber tasks.** Lyptus Research said GPT-5.5 now saturates its dataset, reaching a 5.1-hour time horizon at a 2M-token budget and solving 92.4% of tasks at 50M tokens, beyond 12 hours. The same benchmark line previously measured about 3 hours for Opus 4.6 at 2M tokens and described a doubling trend every six months since 2024; separately, a researcher said GPT-5.5 found a real 27-year-old RCE after checking the commit history [2, 3]. Capability gains are now showing up in both benchmark saturation and real bug-finding.
- **Enterprise IT remains a hard benchmark.** Artificial Analysis and

IBM Research launched ITBench-AA for Kubernetes incident response and found every frontier model below 50% accuracy, led by Claude Opus 4.7 at 47% and GPT-5.5 at 46%. They also found that longer trajectories often hurt: GPT-5.5 averaged 31 turns per task at about 46%, while Gemini 3.1 Pro averaged 83 turns at 30% [4, 5]. Strong general-purpose models still need much better harnesses and workflows for enterprise ops.

Research & Innovation

Why it matters: the best research updates were about cutting training cost, improving self-improvement, and expanding AI into biology.

- **Sakana AI's DiffusionBlocks reframes network training block by block.** The method trains one block at a time, needs memory for only a single block, and matched end-to-end performance across ViT, DiT, masked diffusion, autoregressive transformers, and recurrent-depth transformers. For looped transformers, it can replace BPTT with a single forward pass during training [6].
- **Biohub released Evolutionary Scale Models.** ESM is positioned as an open engine for protein prediction, design, and discovery, with a protein language model, ESMFold2, and an atlas containing 6.8 billion sequences and 1.1 billion predicted structures. The release says it has already designed cancer-related proteins and a PD-L1-binding antibody-like protein that worked in lab tests [7].
- **Self-Verified Distillation offers a lighter path to improvement.** The method lets an already post-trained reasoning model generate answers, verify them itself, and train only on responses that pass verification, without ground-truth answers or external verifiers [8, 9].

Products & Launches

Why it matters: launches focused on practical retrieval, search, and document-processing tools rather than just bigger chatbots.

- **Google DeepMind released Gemini Embedding 2.** It is described as the company's first native multimodal embedding model, creating a unified representation for text, audio, video, and image inputs [10].
- **Surya OCR 2 raised the bar for open OCR.** The 650M-parameter model scored 83.3% on the olmocr benchmark and 87% on an internal 91-language benchmark, with reported gains on tables, handwriting, forms, math, and layout. It runs on CPU, GPU, and MPS, with 5 pages per second on an RTX 5090 [11, 12, 13].
- **Ask YouTube turns video search into a conversation.** Google said the feature handles complex queries, supports follow-up questions, and returns structured responses built from relevant long-form videos and Shorts. It is live for Premium users in the U.S. and rolling out more broadly [14].

Industry Moves

Why it matters: companies are now funding the layers above static models: continual learning, infrastructure, and social transition.

- **Trajectory launched around continual learning.** The startup says it uses product-usage signals to continuously post-train agentic models, has raised $15M, and is already working with companies including Clay, Harvey, Decagon, Mercor, and Rogo [15].
- **Modal raised a $355M Series C** to expand its AI cloud infrastructure platform [16].
- **OpenAI Foundation committed an initial $250M** to measurement, transition support, and new approaches to broadly shared prosperity as AI reshapes work and the economy [17, 18].

Policy & Regulation

Why it matters: even without a major government ruling today, enterprise AI deployment is becoming more compliance-heavy.

- **OpenAI added more governance controls for enterprise use.** Its Admin API now supports spend alerts, model allowlists, data-retention controls, hosted tool controls, and more granular cost visibility; it also added Workload Identity Federation and support for private MCP servers over outbound-only HTTPS [19, 20, 21].

Quick Takes

Why it matters: several smaller releases sharpened the picture on speed, cost, and agent infrastructure.

- **Qwen3.5 on TokenSpeed hit 580 tokens per second** for agentic workloads on NVIDIA GPUs [22].
- **Perplexity open-sourced a Unigram tokenizer** that cuts CPU utilization by 5-6x and runs in 63 microseconds at 514 tokens [23, 24].
- **Deep Agents v0.6 added Delta channels,** cutting one 200-turn coding session's checkpoint storage from 5.3GB to 129MB [25].
- **Claude Code shipped reliability upgrades,** including self-healing sessions plus MCP, streaming, and renderer fixes [26, 27, 28, 29, 30].

Sources

1. X post by @cognition
2. X post by @LyptusResearch
3. X post by @PhiloGroves
4. X post by @ArtificialAnlys
5. X post by @ArtificialAnlys

6. X post by @SakanaAILabs
7. X post by @TheTuringPost
8. X post by @tonyh_lee
9. X post by @percyliang
10. X post by @mseyed
11. X post by @VikParuchuri
12. X post by @VikParuchuri
13. X post by @VikParuchuri
14. X post by @Google
15. X post by @rronak_
16. X post by @StasBekman
17. X post by @sama
18. X post by @woj_zaremba
19. X post by @OpenAIDevs
20. X post by @OpenAIDevs
21. X post by @OpenAIDevs
22. X post by @PyTorch
23. X post by @perplexity_ai
24. X post by @perplexity_ai
25. X post by @LangChain
26. X post by @ClaudeDevs
27. X post by @ClaudeDevs
28. X post by @ClaudeDevs
29. X post by @ClaudeDevs
30. X post by @ClaudeDevs