Real-Time AI Interfaces, OpenAI's Enterprise Push, and Devin's Growth

AI High Signal Digest

2026-05-12

Real-Time AI Interfaces, OpenAI's Enterprise Push, and Devin's Growth

By AI High Signal Digest • May 12, 2026

Thinking Machines pushed AI toward real-time multimodal collaboration, while OpenAI expanded into enterprise deployment and cybersecurity workflows. This brief also covers key research advances, new agent products, and fresh market signals from Anthropic, Cognition, Core Automation, and Cerebras.

Top Stories

Why it matters: The biggest updates point to AI moving beyond turn-based chat and model access toward real-time collaboration, domain-specific workflows, and deeper enterprise deployment.

- **Thinking Machines introduced interaction models.** The system is designed to talk, listen, watch, think, and collaborate simultaneously in real time, with demos showing interruption handling, continuous audio/video processing, and visually proactive tasks like posture monitoring and live finger counting [1, 2, 3, 4]. *Impact:* This is a direct attempt to move AI UX beyond prompt/response into continuous multimodal collaboration.
- **OpenAI widened its enterprise push on two fronts.** It launched the **OpenAI Deployment Company**—majority-owned by OpenAI, backed by 19 partners, starting with 150 forward-deployed engineers and deployment specialists plus $4B of initial investment—and also launched **Daybreak**, which pairs OpenAI models and Codex with security partners to scan repositories, find vulnerabilities, generate patches, and automate response [5, 6, 7, 8, 9]. *Impact:* OpenAI is expanding beyond model access into implementation and security-specific workflows.
- **Cognition's Devin is showing large commercial traction.** In its

first 18 months, Devin reached a **$445M revenue run rate**, with usage doubling every eight weeks; customers include the **US Army, Goldman Sachs, and Mercedes-Benz**, and Cognition is raising at around **$25B** valuation [10]. *Impact:* The software-agent category now has a major revenue datapoint.

Research & Innovation

Why it matters: The most notable technical work focused on more predictable training, alternative language-modeling methods, and efficient small multimodal models.

- **Marin's Delphi** predicted a **25B-parameter, 600B-token** training run by extrapolating **300x** from smaller models, with reported **0.2% error** [11].
- A new paper on **entropy-gated continuous bitstream diffusion** says diffusion over bitstreams can outperform masked and uniform diffusion baselines and essentially match autoregressive language models under the paper's evaluation settings [12].
- **MiniCPM-V 4.6 1.3B Instruct** scored **13** on the Artificial Analysis Intelligence Index—the highest for open weights under 2B parameters—while using just **5.4M** output tokens and reaching **38%** on MMMU-Pro [13].

Products & Launches

Why it matters: New releases kept pushing agents closer to live operations across meetings, coding sessions, and desktop workflows.

- **GPT-Realtime-2** was demoed as a meeting agent that can turn spoken standup updates into ticket moves; OpenAI also released a repo for building similar voice-to-action workflows [14, 15].
- **Claude Code** added **agent view**, a research-preview control plane that shows all sessions in one list; terminal users can manage it via `claude agents` [16, 17].
- **Hermes Agent** previewed **computer use with any model**, letting models control a user's computer in the background while the user keeps keyboard, mouse, and screen control [18, 19].

Industry Moves

Why it matters: Capital and distribution are clustering around enterprise adoption, new labs, and AI infrastructure.

- **Anthropic launched Claude Platform on AWS**, giving AWS customers native Claude access with AWS authentication, billing, commitment retirement, and governance tooling; Anthropic says it is a distribution and enterprise adoption move, not a new model [20, 21].

- **Core Automation**, a six-week-old startup founded by ex-OpenAI researcher **Jerry Tworek**, is already seeking funding at a **$4B valuation**; it is building models that continuously learn from real-world experience, with **Nvidia** as an early backer [22].
- **Cerebras** is reportedly increasing the size and price of its IPO after demand exceeded available shares by **20x** [23].

Quick Takes

Why it matters: These smaller updates still sharpen the picture on benchmarks, infrastructure, defense, and real-world AI usage.

- **Epoch AI** says an AI-assisted review of **FrontierMath** flagged fatal errors in about a third of Tier 1–4 problems; corrected scores will follow after human review [24].
- **vLLM** now tops Artificial Analysis on **DeepSeek V3.2** and says its leading deployments for DeepSeek, MiniMax-M2.5, and Qwen 3.5 397B are open source [25].
- **Sphere Semi** says its AI-designed chip is now deploying into military hardware with **Northrop Grumman**, calling it the first AI-designed semiconductor to go from concept to deployment in a defense system [26].
- A **METR** survey of 349 technical workers found self-reported AI gains of **1.6–2.1x** in work value on average, while explicitly warning those perceptions likely overestimate ground truth [27, 28].

Sources

1. X post by @thinkymachines
2. X post by @johnschulman2
3. X post by @liliyu_lili
4. X post by @liliyu_lili
5. X post by @OpenAI
6. X post by @gdb
7. X post by @OpenAI
8. X post by @OpenAI
9. X post by @TheRundownAI
10. X post by @colossusmag
11. X post by @WilliamBarrHeld
12. X post by @LucaAmb
13. X post by @ArtificialAnlys
14. X post by @OpenAIDevs
15. X post by @OpenAIDevs
16. X post by @claudeai
17. X post by @_catwu
18. X post by @Teknium

19. X post by @Teknium
20. X post by @kimmonismus
21. X post by @kimmonismus
22. X post by @theinformation
23. X post by @kimmonismus
24. X post by @EpochAIResearch
25. X post by @vllm_project
26. X post by @StevenGlinert
27. X post by @METR_Evals
28. X post by @METR_Evals