NVIDIA Opens an Omni Stack, Codex Broadens Its Reach, and Google Signs a Pentagon AI Deal

AI High Signal Digest

2026-04-29

NVIDIA Opens an Omni Stack, Codex Broadens Its Reach, and Google Signs a Pentagon AI Deal

By AI High Signal Digest • April 29, 2026

NVIDIA released an open multimodal model designed for agent workflows, OpenAI expanded Codex well beyond coding, and reports detailed Google's classified Pentagon AI deal. Also in focus: stronger math capability signals, on-device privacy tooling, and major partnerships from Profluent, Exa, and Anthropic.

Top Stories

Why it matters: The biggest signals today were an open multimodal push from NVIDIA, a broader scope for Codex, and stronger evidence that frontier models are contributing to serious technical work.

- **NVIDIA released a new open multimodal model built for agent loops.** Nemotron 3 Nano Omni combines audio, image, video, and text in one reasoning loop, ships with **30B parameters** and **256K context**, and quickly landed across vLLM, Together AI, fal, and Ollama. fal highlighted roughly **9× higher throughput** from fewer inference hops in multimodal agent workflows [1, 2, 3, 4, 5].
- **Codex moved closer to a general work agent.** Recent updates added macOS computer use, an in-app browser for inspecting localhost builds, built-in image generation, plugins, first-class artifacts, and follow-up automations. OpenAI also added a **/fast** mode for GPT-5.5 in Codex at **1.5×** speed and reset rate limits for all paid plans [6, 7, 8, 9, 10, 11, 12, 13, 14].
- **OpenAI's math signal kept strengthening.** OpenAI said **GPT-5.4 Pro** helped solve a **60-year-old Erdős problem**, while **GPT-5.5 Pro** reached a new high of **159** on Epoch's Capabilities Index and improved

FrontierMath results, including solving two previously unsolved Tier 4 problems across runs [15, 16, 17].

Research & Innovation

Why it matters: The most useful research today focused on where current systems still break: retrieval, post-training efficiency, and safety visibility.

- **MathNet exposed a major retrieval gap in math AI.** The MIT benchmark includes **30,676 Olympiad-level problems** from **47 countries** and **17 languages**; top models reached **78.4%** problem-solving accuracy, but retrieval Recall@1 was only about **5%**, with RAG improving results by up to **12%** [18].
- **Self-distillation is emerging as a serious post-training alternative.** MIT and ETH Zurich researchers described a setup where models act as their own teacher using feedback or demonstrations; they highlighted **SDPO** for RL, **SDFT** for continual learning, and argued the approach is simpler and faster than **GRPO**, with production use already underway [19].
- **A new "Introspection Adapter" targets hidden model behavior.** Researchers trained a single adapter that makes finetuned models describe their behavior and generalizes to detecting hidden misalignment, backdoors, and safeguard removal [20, 21].

Products & Launches

Why it matters: The most notable launches were practical: privacy, enterprise research, and deployable coding models.

- **OpenAI shipped Privacy Filter.** It is a **1.5B-parameter**, open-source, on-device model for PII detection and redaction, scored at **96% F1** on PII-Masking-300k, and can detect sensitive text including **API keys** [22, 23, 24].
- **Google launched Deep Research and Deep Research Max.** The new Gemini 3.1 Pro-powered agents combine open-web search with proprietary enterprise data via **MCP** in a single API call [25].
- **Poolside released its first open-weight coding model. Laguna XS.2** is a **33B total / 3B active** MoE for agentic coding and long-horizon tasks, trained in-house, runnable on a single GPU, and released under **Apache 2.0** [26].

Industry Moves

Why it matters: Partnerships are increasingly about distribution, workflow control, and high-value verticals rather than just model access.

- **Profluent signed a major pharma deal with Eli Lilly.** The partnership is worth **$2.25B plus royalties** and focuses on AI-designed proteins for **large gene insertion** therapeutics [27].
- **Google added Exa search inside Gemini.** Exa said its agent-first search now powers **Grounding With Exa** for Gemini, giving models access to billions of websites, technical docs, papers, people, and companies [28].
- **Anthropic pushed Claude deeper into creative software.** New partnerships with **Blender, Autodesk, Adobe, Ableton**, and others connect Claude directly to professional creative workflows; the Blender connector can debug scenes, build tools, and batch-apply changes across objects [29, 30].

Policy & Regulation

Why it matters: Government AI contracts are becoming more consequential for both deployment norms and internal company politics.

- **Google's Pentagon contract became one of the day's biggest governance stories.** Posts citing *The Information* said Google signed a classified deal allowing use of its AI for "any lawful government purpose" and requiring help adjusting safety filters; more than **600 employees** reportedly opposed the move, and lawyers said the contract's "not intended for" language on surveillance and autonomous weapons carries no legal weight [31, 32, 33].

Quick Takes

Why it matters: A few smaller releases still stood out for real-time multimodality, evaluation infrastructure, and world-model tooling.

- **MiniCPM-o 4.5** open-sourced a **9B** full-duplex multimodal streaming model and said it can run offline on Windows and macOS hardware [34].
- **fal** launched **World Model Accelerator**, an inference engine for generative media and world models that scales from **1 to 1,000+ GPUs** [35].
- **ParseBench** launched with **2,000 verified pages** from real enterprise documents plus a Kaggle leaderboard for document understanding [36, 37].
- **VibeBench** is recruiting **1,000** software engineers to rank models on real engineering work, with public reports planned after each evaluation round [38].

Sources

1. X post by @NVIDIAAI
2. X post by @vllm_project
3. X post by @togethercompute
4. X post by @fal
5. X post by @ollama
6. X post by @reach_vb
7. X post by @reach_vb
8. X post by @reach_vb
9. X post by @reach_vb
10. X post by @reach_vb
11. X post by @reach_vb
12. X post by @reach_vb
13. X post by @reach_vb
14. X post by @thsottiaux
15. X post by @OpenAI
16. X post by @EpochAIResearch
17. X post by @EpochAIResearch
18. X post by @TheTuringPost
19. X post by @yacinelearning
20. X post by @kshenoy_
21. X post by @NeelNanda5
22. X post by @dl_weekly
23. X post by @thursdai_pod
24. X post by @thursdai_pod
25. X post by @dl_weekly
26. X post by @poolsideai
27. X post by @thisismadani
28. X post by @ExaAILabs
29. X post by @Techmeme
30. X post by @claudeai
31. X post by @kimmonismus
32. X post by @erinkwoo
33. X post by @TheRundownAI
34. X post by @OpenBMB
35. X post by @fal
36. X post by @osanseviero
37. X post by @jerryjliu0
38. X post by @jpschroeder