OpenAI's $122B Raise, Anthropic's Leak, and a Benchmark Reset for Multimodal AI

AI High Signal Digest

2026-04-01

OpenAI's $122B Raise, Anthropic's Leak, and a Benchmark Reset for Multimodal AI

By AI High Signal Digest • April 1, 2026

This brief covers OpenAI's massive financing and platform push, the Claude Code leak and what it revealed about proactive agents, Stanford's challenge to multimodal benchmarks, and key launches across video, spreadsheets, and enterprise copilots.

Top Stories

Why it matters: This cycle was defined by capital concentration, a rare agent-code leak, a challenge to multimodal benchmark validity, and stronger evidence that useful AI can run much closer to the edge.

OpenAI paired massive financing with a broader product ambition

OpenAI said it closed its latest funding round with **$122 billion** in committed capital at an **$852B post-money valuation** [1]. The company said the funding gives it resources to lead at scale and expand AI's benefits by putting useful intelligence in people's hands early [1]. Separate posts interpreting the announcement framed the next phase as consolidation of ChatGPT, Codex, browsing, and agents into a **single AI superapp** [2, 3]. Widely shared posts also cited steep commercialization progress, including **$1B within a year of ChatGPT**, **$1B per quarter by end-2024**, and **$2B per month now** [4, 2].

Impact: OpenAI is pairing balance-sheet scale with a platform strategy, raising the competitive bar on both infrastructure and distribution.

The Claude Code leak exposed Anthropic's proactive-agent design

Multiple posts said **Claude Code** source code leaked through an npm source map [5]. Reviews of the leaked code described **KAIROS** as an always-on proactive mode behind internal feature flags, with heartbeat prompts, push notifications, file delivery, pull-request subscriptions, append-only daily logs, and nightly memory consolidation via **autoDream** [6]. Posts reviewing the leak also said the code referenced unreleased Anthropic model names and variants including **Mythos/Capybara**, **Opus 4.7**, and **Sonnet 4.8** [7]. Anthropic then sent **DMCA requests** against repositories carrying the leaked code [8, 9], and an official statement on the leak was reported [10].

> "every few seconds, KAIROS gets a heartbeat. basically a prompt that says 'anything worth doing right now?' " [6]

Impact: The leak offered a rare view into how frontier coding agents may move from reactive copilots toward background autonomy, while also highlighting the security and IP fragility of agent products.

Stanford's MIRAGE result challenged multimodal evaluation

A widely shared summary of Stanford's **MIRAGE** paper, co-authored by Fei-Fei Li, said leading vision-language models still scored **70-80%** on six major vision benchmarks even after images were silently removed [11]. The same summary said a **3B** text-only super-guesser trained on text from chest X-ray questions ranked **#1** on held-out tests, beating VLMs and radiologists [11]. A cleanup method called **B-Clean** reportedly removed **74-77%** of questions from existing vision benchmarks because they did not truly test vision [11].

Impact: If these reported results hold up, current multimodal leaderboards may be overstating visual understanding and understating shortcut exploitation—especially in medical settings [11].

PrismML pushed 1-bit local models into the spotlight

PrismML emerged from stealth arguing that the next AI gains will come from **intelligence density** rather than only parameter count [12]. Its **1-bit Bonsai 8B** model fits in **1.15GB** of memory and is described as **14x smaller**, **8x faster**, **5x more energy efficient**, and over **10x** the intelligence density of its full-precision counterparts, while remaining competitive in its class; Bonsai **8B**, **4B**, and **1.7B** were open-sourced under Apache 2.0 [12]. PrismML says this should enable **on-device agents**, **real-time robotics**, and **offline intelligence** [12]. A follow-up post said the 1-bit Bonsai family shifts the Pareto frontier of **intelligence vs. size** dramatically to the left [13], and a demo showed Bonsai 8B running locally on an M4 Pro with much lower memory use and higher throughput than a standard 16-bit 8B model [14].

Impact: Small local models are starting to look less like a fallback and more like a distinct product and infrastructure strategy.

Research & Innovation

Why it matters: The most interesting technical work this cycle focused on better reasoning training, longer-lived agent memory, smaller useful models, and more reliable evaluation.

- **OpenAI on Erdős problems:** OpenAI researchers said an internal model found **short and elegant** proofs for **three further open problems** due to Erdős, with the paper posted on arXiv [15]. A separate OpenAI executive post framed the broader trend as AI solving more open problems while producing more elegant proofs as models improve [16].
- **Token-level RL credit assignment:** Qwen Pilot introduced **FIPO**, which uses a GAE-style Future KL signal to assign credit to individual tokens during reasoning. The claim is that, unlike GRPO, it can reinforce helpful tokens and suppress derailing ones, producing longer and more accurate chains beyond **10k tokens** with strong gains on **AIME24** [17].
- **Long-term memory for agents: GAAMA** proposes a hierarchical memory system that combines RAG with knowledge graphs. The reported result is **78.9% mean reward** on **LoCoMo-10**, outperforming HippoRAG and tuned RAG baselines [18]. The core claim is that graph-augmented retrieval plus higher-order reflections improves multi-session recall [18].
- **Useful small models kept improving:** Liquid AI released **LFM2.5-350M**, a **350M-parameter** model aimed at **agentic loops**, reliable data extraction, and tool use [19]. It was trained on **28T tokens** with scaled RL [20], with reported gains from LFM2-350M in instruction following (**18.20 → 40.69**), data extraction (**11.67 → 32.45**), and tool use (**22.95 → 44.11**) [20]. Quantized size is under **500MB**, making it usable in constrained environments [19].
- **GPU kernel scheduling got more automated:** Modular said it built a **constraint solver in Mojo** that automatically derives pipeline schedules for GPU kernels, tackling the complexity of FA4 on Blackwell with **14 ops**, **5 hardware units**, and **28 dependency edges** [21]. The reported outcome is simpler kernels, race conditions defined away, and more portable intra-kernel composition while keeping full hardware control [22, 23].
- **Benchmark methodology is getting more careful:** Google Research announced a new framework for improving benchmark reproducibility by optimizing the ratio of items to human raters per item, with the goal of better capturing human disagreement in subjective tasks [24].

Products & Launches

Why it matters: Vendors are turning multi-model orchestration, cheaper video generation, spreadsheet workflows, and agent interfaces into products people can actually use.

- **Microsoft pushed multi-model workflows into M365 Copilot: Council** lets users run multiple models on the same prompt to compare where they align and diverge [25]. **Critique** is a new multi-model deep research system that Microsoft says uses multiple models together to generate better responses and reports, with a feedback loop aimed at improving factual accuracy, analytical breadth, and presentation [26, 27].
- **Veo 3.1 Lite widened access to video generation:** Google made **Veo 3.1 Lite** available in the **Gemini API** and **Google AI Studio** for rapid prototyping and high-volume generation at **$0.05/sec**, or half the cost of Veo 3.1 Fast [28]. It supports **text-to-video** and **image-to-video**, **16:9** and **9:16** output, and **4s, 6s, and 8s** clips [28]. Fal.ai also put Veo 3.1 Lite live with **first-last-frame-to-video** and both **720p** and **1080p** options [29, 30].
- **OpenAI expanded practical workflow surfaces: ChatGPT for Excel** is now available worldwide except EU consumer plans [31, 32]. Separately, the **GitHub plugin in the Codex app** can review issues, address feedback, commit changes, and open pull requests [33].
- **Google AI Studio added music tooling: Music Playground**, powered by **Lyria 3**, launched with a **Composer Mode** that lets users describe music, hear it, then export the result to code and build from it [34].
- **Agent interfaces kept broadening:** Perceptron launched an **MCP server** that gives agents stronger vision via Isaac at lower cost than general-purpose multimodal models [35, 36]. In open-source tooling, a new Hermes Agent PR added **computer use** on a real Mac from a phone, with no sandbox and real-time control over desktop apps [37].

Industry Moves

Why it matters: Companies are reorganizing around agents, security, and open-model infrastructure rather than treating AI as an isolated feature.

- **OpenAI broadened its infrastructure posture:** A reported partnership with **Amazon** would build infrastructure for **AI agents** on AWS, signaling a wider cloud posture around deployment [38].
- **Microsoft formalized its OpenClaw bet:** Omar Shahine said he joined Microsoft to bring **OpenClaw + personal agents** to **Microsoft 365**, with a goal of proactive workplace assistants that take on tasks end-to-end; he also said a fully integrated **Teams plugin** is already deployed [39].
- **Perplexity moved into security research:** The company launched the **Secure Intelligence Institute**, led by Purdue's **Dr. Ninghui Li**, to work with top cryptography, security, and ML teams [40]. Its first paper responds to **NIST's** request for information on securing autonomous agents [41].
- **Open-model enterprise adoption kept strengthening:** Hugging

Face CEO Clement Delangue said companies including **Pinterest, Airbnb, Notion, Cursor, and Intercom** are finding it **better, cheaper, faster** to use and train open models in-house for many tasks [42]. Hugging Face also released **TRL v1** with **75+** post-training methods including SFT, DPO, GRPO, and async RL [43].
- **QodoAI raised more capital for AI coding infrastructure:** QodoAI announced a **$70M** raise, with the company arguing that software development has fundamentally changed but that enterprise-grade transformation is still early [44].
- **Gemma's ecosystem scale kept growing:** Two years after launch, Google's **Gemma** family of open models reached **400M downloads** and **100,000 variants** [45].

Policy & Regulation

Why it matters: Formal regulation remains uneven, but the policy surface is expanding through safety partnerships, legislative proposals, legal enforcement, and geopolitical risk.

- **Australia and Anthropic signed a safety MOU:** Anthropic said it signed an **MOU** with the **Australian Government** to collaborate on **AI safety research** and support Australia's **National AI Plan** [46].
- **US debate over AI rules intensified:** Sen. **Bernie Sanders** said **74%** of Americans believe the government is not doing enough to regulate AI and pointed to his proposed **moratorium bill** as a way to address AI risks and broaden who benefits [47]. Separately, **Andrew Ng** said he supports the White House's **proposed** national AI legislative framework with federal preemption to avoid a patchwork of state-level restrictions [48].
- **Anthropic's leak response turned legal:** After the Claude Code leak, Anthropic sent **DMCA requests** to shut down repositories hosting the source code [8, 9].
- **Geopolitical risk to AI infrastructure rose:** A cited post reported that the IRGC accused American AI companies of being 'the primary element in designing and tracking assassination targets' and threatened to treat them as 'legitimate targets' [49]. Another post interpreted that as a threat to data centers [49].

Quick Takes

Why it matters: These smaller signals help track where capability, adoption, and risk are moving next.

- **KAT-Coder-Pro V2** reached **44** on the Artificial Analysis Intelligence Index, matching Claude Sonnet 4.6 among non-reasoning models. Reported strengths were **49%** on Terminal-Bench Hard, about **109** output tokens/sec, and **$73** benchmark cost; reported weaknesses were long-context reasoning and knowledge regressions versus V1 [50].

- **IBM Granite 4.0-3B-Vision** launched as a document-focused VLM with state-of-the-art performance for its size on tables and charts, compatibility with Transformers and vLLM, and a free license [51].
- **Qdrant Agent Skills** positions vector search as structured, composable retrieval for agents. Qdrant's reported comparison showed **96% vs 65%** pass rate, **1.8x** faster execution, **13%** fewer tokens, and **3x** more consistency with Skills enabled [52].
- **OpenRouter's Model Fusion** combines outputs from multiple models into one answer; OpenRouter said every Deep Research agent preferred the fused response over its own in testing, and the feature does not require a subscription [53].
- **LangChain** added more operational guidance for teams putting agents into production, including a free course on **monitoring production agents** and a trace-centered **agent improvement loop** guide built around costs, latency, evals, prompt injection, and PII leakage [54, 55].
- **Arena rankings kept shifting: Claude Opus 4.6** stayed on top of Text Arena, while **Gemini-3.1 Pro**, **GPT-5.4 High**, and **Grok-4.20 (Reasoning)** entered the top 10 [56]. Grok-4.20 also landed **#3** in Medicine & Healthcare and **#6** across Expert Prompts, Math, and Legal & Government slices [57].
- **Security risk in the AI developer stack stayed elevated:** A security roundup said TeamPCP poisoned tools including **LiteLLM**, the **axios** npm incident gave attackers remote control on affected machines, and AI-software pace may be amplifying classic supply-chain failures and human error [58].

Sources

1. X post by @OpenAI
2. X post by @TheRundownAI
3. X post by @kimmonismus
4. X post by @reach_vb
5. X post by @Fried_rice
6. X post by @itsolelehmann
7. X post by @kimmonismus
8. X post by @dbreunig
9. X post by @BlancheMinerva
10. X post by @theo
11. X post by @heygurisingh
12. X post by @PrismML
13. X post by @PrismML
14. X post by @PrismML
15. X post by @mehtaab_sawhney

16. X post by @kevinweil
17. X post by @sheriyuo
18. X post by @dair_ai
19. X post by @liquidai
20. X post by @liquidai
21. X post by @Modular
22. X post by @clattner_llvm
23. X post by @clattner_llvm
24. X post by @GoogleResearch
25. X post by @satyanadella
26. X post by @satyanadella
27. X post by @yusuf_i_mehdi
28. X post by @_philschmid
29. X post by @fal
30. X post by @fal
31. X post by @ryanbrewer
32. X post by @snsf
33. X post by @OpenAIDevs
34. X post by @GoogleAIStudio
35. X post by @perceptroninc
36. X post by @AkshatS07
37. X post by @0xbyt4
38. X post by @DeepLearningAI
39. X post by @OmarShahine
40. X post by @perplexity_ai
41. X post by @perplexity_ai
42. X post by @ClementDelangue
43. X post by @ClementDelangue
44. X post by @itamar_mar
45. X post by @osanseviero
46. X post by @AnthropicAI
47. X post by @SenSanders
48. X post by @AndrewYNg
49. X post by @peterwildeford
50. X post by @ArtificialAnlys
51. X post by @mervenoyann
52. X post by @qdrant_engine
53. X post by @OpenRouter
54. X post by @LangChain
55. X post by @LangChain
56. X post by @arena
57. X post by @arena
58. X post by @saranormous