Anthropic Leak, Compute Bottlenecks, and the Agent Playbook Take Center Stage

AI High Signal Digest

2026-03-28

Anthropic Leak, Compute Bottlenecks, and the Agent Playbook Take Center Stage

By AI High Signal Digest • March 28, 2026

The brief covers leaked Anthropic model details and the security fallout, tightening memory and power bottlenecks, the steady open-vs-closed model gap, and new research and product launches across agents, voice, vision, and chip design.

Top Stories

Why it matters: Four themes stood out: frontier-model security, physical infrastructure constraints, the economics of open vs. closed models, and a more formal operating model for AI agents.

Anthropic's unreleased model leak became a security story

According to posts citing leaked materials, Anthropic has been testing a model called *Mythos* with select customers. Those posts described it as a new tier above Opus—later edited in one post to *Capybara*—with stronger results in coding, academic reasoning, and cybersecurity, plus a slow rollout because of compute intensity and security concerns [1, 2, 3]. Fortune was separately cited for reporting that Anthropic left details of an unreleased model in an unsecured data trove [4].

Impact: Frontier-model competition is now tied not just to capability, but to selective access, cyber risk, and operational security [5, 4].

Compute constraints are showing up in memory, power, and construction schedules

Epoch AI said the total memory bandwidth of AI chips shipped since 2022 has reached 70 million terabytes per second and is growing 4.1x per year, while AI

inference is often bottlenecked by memory bandwidth rather than raw compute [6, 7]. It also said AI chips consumed more than 90% of total HBM production in 2025 and that HBM prices spiked in early 2026 as demand outpaced supply [8]. At the same time, Microsoft said it is partnering with Crusoe on a 900MW AI factory in Abilene, Texas [9], OpenAI said steel beams went up this week at its Michigan Stargate site with Oracle and Related Digital [10], and NVIDIA said Vera Rubin + Groq 3 LPX can deliver up to 35x more performance per megawatt for trillion-parameter models and massive context workloads [11].

Impact: The competitive bottleneck is increasingly about watts, memory bandwidth, and buildout speed—not only model quality [7, 8, 9, 10].

The open/closed gap is much smaller than it used to be, but the frontier is still closed

Arena said the gap between top open-source and proprietary text models has held at roughly 50-60 points for about 14 months, down from 100-150 points before mid-2024 [12]. It also said proprietary models currently occupy the first 20 places on the Text Arena leaderboard, while the leading open models are GLM-5 at #20, Kimi-K2.5-Thinking at #23, and Qwen3.5-397b-a17b at #27 [12]. In separate Arena analysis, GPT-5.4 High, Mini, and Nano behaved like scaled versions of the same model, suggesting price differences mainly reflect efficiency rather than different core capabilities [13].

Impact: Open models are closer than before, but the leading edge still sits with closed labs, and pricing is becoming more about efficiency per task than a simple proxy for intelligence [12, 13].

The agent era is getting its own playbook

A new Google-linked report argues that intelligence explosions are social rather than individual, and that future progress may come from *human-AI configurations and agent institutions* rather than bigger monolithic models [14]. In plain language, the argument is that groups of agents with roles, checks, and protocols may matter more than one ever-larger model [14].

> Every prior intelligence explosion in human history was social, not individual. [14]

IBM's new survey on workflow optimization for LLM agents organizes agent systems by when workflow structure is set, what components are optimized, and which signals guide the optimization [15]. Artificial Analysis also launched AA-AgentPerf, a hardware benchmark for the agent era that uses real coding-agent workloads and reports maximum concurrent users per accelerator, per kW, per dollar, and per rack [16, 17].

Impact: The discussion is moving from which single model is best to how agent systems should be structured, evaluated, and deployed [15, 16].

Research & Innovation

Why it matters: Research attention is shifting toward unified multimodal systems, better long-context reasoning, more stable world models, and more realistic evaluations.

- **Apple AToken:** Apple introduced AToken, a shared tokenizer and encoder for images, video, and 3D objects in one framework. The post said it beats or rivals specialized models and allows knowledge transfer across media types [18].
- **SAGE:** This closed-loop multi-agent training method co-evolves a Challenger, Planner, Solver, and Critic from one LLM backbone using just 500 seed examples. On Qwen-2.5-7B, it reportedly improved out-of-distribution performance by 4.2% while maintaining in-distribution accuracy [19].
- **Together Research's divide-and-conquer approach:** A Planner rewrites tasks for parallel Workers and a Manager combines their outputs. Together said Llama-3-70B and Qwen-72B using this setup can match or beat GPT-4o single-shot on long-context retrieval, QA, and summarization as context length grows, though the method still struggles when important clues are spread across distant chunks [20, 21, 22, 23].
- **LeWorldModel:** Yann LeCun's team released LeWorldModel, described as a world model that avoids collapse by adding a SIGReg regularizer to its prediction loss. The post also claimed 15M parameters, training on one GPU in hours, 48x faster planning, and about 200x fewer tokens for encoding [24].
- **CursorBench:** A new benchmark for coding agents uses real Cursor team coding sessions, evaluates more than functional correctness, emphasizes long-horizon tasks with a median 181 lines changed per task, and keeps the data refreshed with recent sessions [25].

Products & Launches

Why it matters: Product releases this cycle focused on deployability: lower-latency voice agents, faster video processing, more local execution, and tools that slot directly into agent workflows.

- **OpenAI gpt-realtime-1.5:** OpenAI showed a clinic concierge demo for a Singapore health clinic. It speaks naturally with patients, collects the needed details, and books appointments in real time [26].
- **Meta SAM 3.1:** Meta released SAM 3.1 as a drop-in update to SAM 3. Its core change is object multiplexing, which lets the model track up to 16 objects in one forward pass and doubles throughput from 16 to 32 FPS on a single H100 for medium-object videos [27, 28]. Meta said the point is to make high-performance video applications feasible on smaller, more accessible hardware [27].
- **Cohere Transcribe in the browser:** Cohere's multilingual speech

recognition model can run entirely locally in a browser on WebGPU. A post said it can transcribe 1 hour of audio in 100 seconds, is fully private, free, and requires no installation [29].

- **LiteParse:** LlamaIndex's LiteParse is a model-free, open-source document parser for AI agents. It processes about 500 pages in 2 seconds on commodity hardware, supports 50+ file formats, and is designed to plug into agent tools, while the authors note it is not meant to replace OCR-heavy workflows for scanned documents [30, 31].
- **Hermes Agent + Hugging Face:** Hermes Agent is positioned as an open-source agent that remembers what it learns through a multi-level memory system and persistent machine access [32]. Hugging Face is now a first-class inference provider inside Hermes, with 28 curated models in the picker and custom access to 100+ more [33, 34].
- **Gemini video creation:** Google added a Create video workflow in Gemini's app and web experience, where users select the tool, describe the video, optionally upload a reference image or choose a template, and generate directly from the interface [35].

Industry Moves

Why it matters: Business activity keeps pointing to three battlegrounds: capital markets, distribution, and AI-shaped hardware.

- **Anthropic IPO talk is getting more concrete:** A post citing reporting said Anthropic is eyeing a Q4 2026 IPO with a raise above $60 billion, that its annualized revenue more than doubled to $19 billion in the first two months of 2026, and that bankers think it could reach public markets before OpenAI because of its enterprise and developer focus plus a shorter projected path to profitability [36].
- **Perplexity expanded Samsung distribution:** Perplexity said it now powers Samsung's Browsing Assist in Samsung Browser on Galaxy Android and Windows [37]. In a separate post, Aravind Srinivas said the broader partnership now reaches a browser pre-installed on more than 1 billion Samsung devices, extends prior work with Bixby, and includes pre-loading on Galaxy S26 devices alongside Gemini [38].
- **Microsoft added more physical capacity:** Mustafa Suleyman said Microsoft is partnering with Crusoe on a 900MW AI factory in Abilene, Texas to add capacity to its AI fleet and support Microsoft AI infrastructure [9].
- **RicursiveAI is betting RL can compress chip design cycles:** Lightspeed said it led RicursiveAI's $300 million Series A in January. The company says its reinforcement-learning-based semiconductor design platform can compress chip development from years to weeks [39].

Policy & Regulation

Why it matters: Formal AI policy is still uneven, but courts, safety packs, and billing controls are increasingly shaping how models are deployed.

- **Anthropic won a major preliminary court ruling:** A federal judge in California indefinitely blocked the Pentagon's effort to label Anthropic a supply chain risk, though the ruling is temporary and a parallel case is still underway in Washington, D.C. [40].
- **OpenAI published a teen safety policy pack:** OpenAI released a set of prompt-based safety policies intended to create age-appropriate protections for teens, and published the repository publicly [41].
- **Gemini API billing is getting harder to overspend:** Starting April 1, Gemini API billing tiers get a monthly spending cap, with API access pausing until the next month or a tier upgrade if the cap is hit. Users can also set per-project spend caps in AI Studio [42].

Quick Takes

Why it matters: These are smaller updates, but they show where tooling, benchmarks, and open-source ecosystems are moving next.

- OpenAI launched a **Codex use-case gallery** with starter prompts that can open directly in the app, and separately reset Codex usage limits across all plans so users can experiment with newly launched plugins [43, 44].
- **GLM-5.1** is now available to all GLM Coding Plan users, and a separate post said **GLM-5.1 will be open source** [45, 46].
- Epoch AI removed one **FrontierMath: Open Problems** item after GPT-5.2 Pro solved it, because the problem did not meet the benchmark's minimum notability bar; it also updated sourcing guidelines afterward [47, 48, 49].
- Hugging Face's **HF Papers CLI** adds semantic search and markdown retrieval for arXiv papers, aimed at supporting autoresearch workflows [50, 51].
- **Strix** packages multi-agent application pentesting with a built-in browser, proxy, terminal, and Python runtime, aiming to cut automated pentesting from weeks to hours [52].
- **React Native ExecuTorch v0.8.0** adds Vision Camera integration for real-time computer-vision inference on live camera feeds, including support for RF-DETR and Liquid AI's vision-language models [53].
- Qdrant is pushing **sparse embeddings** for e-commerce search, arguing they preserve exact matches and interpretability better than dense embeddings for product attributes such as SKU, size, and brand [54].
- Huawei's **950PR** AI chip was priced at ¥70,000 with a 2H shipment target of 750,000 units, while one commenter argued it is not comparable to Nvidia's H200 for training workloads [55, 56].

Sources

1. X post by @deredleritt3r
2. X post by @kimmonismus
3. X post by @kimmonismus
4. X post by @jeremyakahn
5. X post by @kimmonismus
6. X post by @EpochAIResearch
7. X post by @EpochAIResearch
8. X post by @EpochAIResearch
9. X post by @mustafasuleyman
10. X post by @sama
11. X post by @NVIDIADC
12. X post by @arena
13. X post by @arena
14. X post by @omarsar0
15. X post by @omarsar0
16. X post by @ArtificialAnlys
17. X post by @ArtificialAnlys
18. X post by @DeepLearningAI
19. X post by @dair_ai
20. X post by @togethercompute
21. X post by @togethercompute
22. X post by @togethercompute
23. X post by @togethercompute
24. X post by @LiorOnAI
25. X post by @cwolferesearch
26. X post by @OpenAIDevs
27. X post by @AIatMeta
28. X post by @AIatMeta
29. X post by @xenovacom
30. X post by @jerryjliu0
31. X post by @jerryjliu0
32. X post by @ClementDelangue
33. X post by @NousResearch
34. X post by @Teknium
35. X post by @GeminiApp
36. X post by @kimmonismus
37. X post by @perplexity_ai
38. X post by @AravSrinivas
39. X post by @lightspeedvp
40. X post by @kimmonismus
41. X post by @dl_weekly
42. X post by @_philschmid

43. X post by @romainhuet
44. X post by @thsottiaux
45. X post by @Zai_org
46. X post by @ZixuanLi_
47. X post by @EpochAIResearch
48. X post by @EpochAIResearch
49. X post by @EpochAIResearch
50. X post by @_akhaliq
51. X post by @_akhaliq
52. X post by @TheTuringPost
53. X post by @swmansion
54. X post by @qdrant_engine
55. X post by @GlennLuk
56. X post by @teortaxesTex