GPT-5.4 “extreme reasoning” rumors, Knuth’s Claude paper, and accelerating agent-native software workflows

AI High Signal Digest

2026-03-05

GPT-5.4 “extreme reasoning” rumors, Knuth’s Claude paper, and accelerating agent-native software workflows

By AI High Signal Digest • March 5, 2026

GPT-5.4 is widely rumored to be close, with reports of a 1M-token context window and a new “extreme reasoning” mode. Also: Donald Knuth publishes “Claude’s Cycles” after Claude Opus 4.6 helps resolve an open problem, and agent tooling accelerates (Codex on Windows, Symphony orchestration, and large enterprise deployments).

Top Stories

1) OpenAI’s GPT-5.4 signals a push toward long-horizon “reasoning modes” + 1M context

Why it matters: If these claims hold, GPT-5.4’s combination of **very long context** and an **“extreme reasoning”** setting points to models designed to run **hours-long** workflows and agent loops, not just chat completion.

What’s being reported:

- GPT-5.4 is expected to ship with an **“extreme” reasoning mode** [1][2], described as using more compute for deeper thinking [3].

[1] post by @steph_palazzolo
[2] post by @kimmonismus
[3] post by @kimmonismus

- Multiple posts cite a ~**1M token context window** (reported as up from **400K** in GPT-5.2) [4][5].
- The Information-linked summaries describe better **long-horizon tasks**, improved memory across multi-step workflows, and use for scientific/complex problems [6][7][8].
- Observers report GPT-5.4 has **"landed on the Arena"**, with uncertainty about which variant is live [9][10]. Some posts suggest a release is "very likely" Thursday [11], while another claims it's "confirmed for today" [12].

Source links shared via X posts: - The Information link (via Techmeme): https://www.theinformation.com/articles/openais-next-ai-model-will-extreme-reasoning [13]

2) Donald Knuth credits Claude Opus 4.6 with solving an open problem—then formalizes the proof

Why it matters: This is a concrete example (from a highly respected computer scientist) of an LLM contributing to new mathematical progress—paired with a human-written formal proof.

- Donald Knuth published a paper titled **"Claude's Cycles"** after **Claude Opus 4.6** solved an open **graph decomposition conjecture** he'd been working on for weeks [14].
- Knuth describes **31 explorations** taking ~**1 hour**, after which he read the output, wrote the formal proof, and concluded: "It seems I'll have to revise my opinions about generative AI one of these days." [15]
- Another post summarizes Knuth saying Claude Opus 4.6 cracked a long-standing **Hamiltonian-cycle conjecture for all odd sizes**, calling it "a joy" to see solved [16].

Paper link (as shared): https://cs.stanford.edu/~knuth/papers/claude-cycles.pdf [17]

---

[4] post by @Techmeme
[5] post by @kimmonismus
[6] post by @kimmonismus
[7] post by @kimmonismus
[8] post by @kimmonismus
[9] post by @legit_api
[10] post by @legit_api
[11] post by @kimmonismus
[12] post by @kimmonismus
[13] post by @Techmeme
[14] post by @BoWang87
[15] post by @BoWang87
[16] post by @slow_developer
[17] post by @BoWang87

3) AI and national security: claims of operational use + procurement controversy intensify

Why it matters: This is the part of the AI landscape where model capability, governance, and accountability collide—often with limited public visibility.

- The Washington Post reports that to strike **1,000 targets in 24 hours in Iran**, the U.S. military used its "most advanced AI" in warfare: **Anthropic's Claude** partnered with the military's **Maven Smart System**, suggesting targets and issuing precise location coordinates [18].
- AI ethicist @mmitchell_ai reacted:

  "People are being killed based (in part) on LLM outputs… Are there people being saved based on LLM outputs?" [19]

- Separately, multiple posts describe a memo attributed to Anthropic CEO **Dario Amodei** calling OpenAI's Pentagon/DoD deal **"safety theater"** and expressing skepticism about OpenAI's safeguards [20][21].
- One thread claims the memo describes Palantir pitching a "classifier" approach for red-line violations and characterizes monitoring as "maybe 20% real and 80% safety theater," with **Anthropic rejecting** and **OpenAI accepting** the package [22].
- The Financial Times is cited by multiple accounts as saying Anthropic leadership is **back in talks with the Pentagon** about an AI deal [23][24].
- A separate post says the U.S. State Department is switching its **'StateChat'** from Claude to **GPT 4.1** [25].

4) Software shifts further toward "agent-native" development: Windows sandboxes + ticket orchestration + enterprise rollouts

Why it matters: Tooling is moving from "assistive coding" to **operational systems** that can run tasks, manage environments, and integrate into enterprise workflows.

- **Codex app on Windows:** OpenAI announced the Codex app is now on Windows with a **Windows-native agent sandbox** and support for Windows developer environments in **PowerShell** [26][27]. The sandbox uses OS-level controls (restricted tokens, filesystem ACLs, dedicated sandbox users) for safer execution [28], and OpenAI provides an open-source

[18] post by @washingtonpost
[19] post by @mmitchell_ai
[20] post by @steph_palazzolo
[21] post by @steph_palazzolo
[22] post by @ns123abc
[23] post by @DeItaone
[24] post by @spectatorindex
[25] post by @TheRundownAI
[26] post by @OpenAIDevs
[27] post by @OpenAIDevs
[28] post by @OpenAIDevs

implementation: https://github.com/openai/codex/tree/main/codex-rs/windows-sandbox-rs [29].

- A separate post highlights that the Windows sandbox is **fully open source** and encourages users to fork/build on it [30][31][32].
- **OpenAI Symphony (orchestration):** described as an orchestration layer that polls project boards and spawns agents for each ticket lifecycle stage [33]. A deeper walkthrough claims it can pull real Linear issues, create fresh workspaces per issue, and keep Codex running until tasks are done [34][35].
- **Enterprise deployment:** Factory says it is partnering with **EY** to enable **more than 10,000 engineers** to ship production-grade software with autonomous agents ("Droids") [36]. Factory positions this as one of the largest enterprise deployments of autonomous dev agents to date [37], with EY reportedly throttling traffic due to rapid adoption [38].

5) DoubleAI claims "autonomous expert" gains in GPU kernel engineering (cuGraph)

Why it matters: GPU kernel optimization is typically a scarce-expertise domain; progress here can compound across the AI stack by improving the performance ceiling of widely used libraries.

- DoubleAI's **WarpSpeed** is claimed to have autonomously rewritten and re-optimized kernels in **cuGraph** across **A100, L4, and A10G** [39].
- Reported results: **3.6× average speedup**, **100% of kernels** improved, and **55%** seeing **>2×** improvement [40]. The hyper-optimized version is released on GitHub as a drop-in replacement (no code changes required) [41].
- The authors frame the work as requiring new algorithmic ideas (Diligent framework, PAC-reasoning, agentic search) for domains with scarce data, hard validation, and long decision chains [42].

[29] post by @OpenAIDevs
[30] post by @reach_vb
[31] post by @reach_vb
[32] post by @reach_vb
[33] post by @scaling01
[34] post by @kevinkern
[35] post by @kevinkern
[36] post by @FactoryAI
[37] post by @FactoryAI
[38] post by @matanSF
[39] post by @AmnonShashua
[40] post by @AmnonShashua
[41] post by @AmnonShashua
[42] post by @AmnonShashua

Research & Innovation

Why it matters: Several releases this cycle target the "hard middle" of AI progress: inference efficiency, multimodal training without brittle dependencies, and agent reliability (memory, evaluation, and proactive interaction).

Multimodal training: Self-Flow (Black Forest Labs)

- Black Forest Labs previewed **Self-Flow**, a scalable approach for end-to-end multimodal generative training across **image, video, audio, text** using self-supervised flow matching (without relying on external pretrained representation models) [43].
- Reported results include **up to 2.8× faster convergence** across modalities, improved temporal consistency in video, and sharper text rendering/typography [44]. BFL frames it as foundational for multimodal visual intelligence [45].
- Additional details shared: it combines per-timestep flow matching with dual-timestep representation learning and is presented as outperforming prior methods with promising scaling behavior [46][47].
- Resources: https://bfl.ai/research/self-flow and code at https://github.com/black-forest-labs/Self-Flow [48].

Teaching models to update beliefs: "reason like Bayesians" (Google Research)

- Google Research introduced a method to teach LLMs to "reason like Bayesians" by training them to mimic **optimal probabilistic inference**, improving their ability to **update predictions** and **generalize across new domains** [49].
- An example task described is a flight recommendation assistant that receives user feedback each round on whether it chose correctly and what the correct answer was [50].

Faster inference: Speculative Speculative Decoding (Together AI)

- Together AI researchers announced **Speculative Speculative Decoding (SSD)**, an inference algorithm reported as **up to 2× faster** than the strongest inference engines [51].

[43] post by @bfl_ml
[44] post by @bfl_ml
[45] post by @bfl_ml
[46] post by @robrombach
[47] post by @robrombach
[48] post by @robrombach
[49] post by @GoogleResearch
[50] post by @_arohan_
[51] post by @tanishqkumar07

- A collaborator notes it applies "asynchronous machine" principles familiar from GPU kernels to speculative decoding [52].

Agent memory: retrieval beats "fancy writing"

- New research introduces a diagnostic framework separating **retrieval failures** from **utilization failures** in agent memory systems [53].
- Core claim: retrieval approach matters far more than writing strategy—accuracy varies ~**20 percentage points** across retrieval methods vs **3–8 points** across writing strategies [54].
- Simple **raw chunking** is reported to match or outperform more expensive alternatives like Mem0-style fact extraction or MemGPT-style summarization [55].
- Paper link: https://arxiv.org/abs/2603.02473 [56].

Proactive agents with implicit human state: NeuroSkill (MIT)

- NeuroSkill is presented as a real-time agentic system integrating **Brain-Computer Interface signals** with foundation models to model human cognitive/emotional state, running fully **offline on the edge** [5758].
- Its **NeuroLoop** harness is described as enabling proactive workflows that respond to both explicit and implicit requests through tool calls [5960].
- Paper: https://arxiv.org/abs/2603.03212 [61].

Biology: open biological models + whole-tissue recording approaches

- The Arc Institute announced **Evo 2**, described as the largest fully open biological AI model to date, published in *Nature* [62]. Goodfire AI says it used interpretability tools to discover "numerous biologically relevant features" in Evo 2 [63].
- A separate Nature paper summary describes **GEMINI** (Granularly Expanding Memory for Intracellular Narrative Integration) as a cellular recorder that encodes activity history in fluorescent "tree-ring" patterns with ~15-minute resolution [64]. AI-based decoding tools are described as central to reading GEMINI's output at whole-brain scale [65].

[52] post by @tri_dao
[53] post by @dair_ai
[54] post by @dair_ai
[55] post by @dair_ai
[56] post by @dair_ai
[57] post by @omarsar0
[58] post by @omarsar0
[59] post by @omarsar0
[60] post by @omarsar0
[61] post by @omarsar0
[62] post by @arcinstitute
[63] post by @GoodfireAI
[64] post by @BoWang87
[65] post by @BoWang87

Products & Launches

Why it matters: The product layer is converging on agent workflows: long-running tasks, memory, multimodal generation, and "do things" interfaces (voice, browser, sandboxes).

Dev + agent tooling

- **Prism + Codex:** OpenAI integrated the Codex harness into Prism (LaTeX environment) to write/compute/analyze/iterate in one place, and added version management [66][67][68].
- **VS Code agents:** VS Code's latest release highlights improved agent orchestration, extensibility, and continuity, including hooks, message steering/queueing, an agentic integrated browser, and shared memory [69][70]. VS Code also notes it will shift from monthly to **weekly** shipments of `main` starting next week [71].
- **Cursor in JetBrains:** Cursor is now available in JetBrains IDEs via the Agent Client Protocol [72].

Research + evidence workflows

- **Elicit API (preview):** Elicit launched an API preview for Pro and Teams users to search **138M+ papers** and generate research reports from code or tools [73]. Docs: http://docs.elicit.com [74]. Examples: https://github.com/elicit/api-examples [75].

Consumer/knowledge tools

- **NotebookLM Studio:** introduced "Cinematic Video Overviews," described as bespoke immersive videos from user sources, rolling out to Ultra users in English [76][77].
- **Google Search (AI Mode) Canvas:** Google is making Canvas in AI Mode available to everyone in the U.S. in English; it supports multi-session planning in a side panel and adds creative writing and coding tasks [78].

[66] post by @vicapow
[67] post by @kevinweil
[68] post by @kevinweil
[69] post by @code
[70] post by @code
[71] post by @pierceboggan
[72] post by @cursor_ai
[73] post by @elicitorg
[74] post by @elicitorg
[75] post by @elicitorg
[76] post by @NotebookLM
[77] post by @NotebookLM
[78] post by @Google

Voice + action interfaces

- **Perplexity Computer Voice Mode:** Perplexity introduced Voice Mode in Perplexity Computer so users can "just talk and do things" [79][80]. The CEO framed it as "Building a kind of JARVIS" [81].

Generative media

- **Kling 3.0 rollout:** Kling AI says Kling 3.0 / Omni / Motion Control are fully rolled out, with features like mocap-level motion control and multi-shot video generation up to 15s [82][83][84]. A creator thread highlights improved micro-expressions and dialogue shots (a "chamber play" becoming easier) [85].
- **Qwen-Image-2.0:** Alibaba introduced Qwen-Image-2.0 with claims including professional typography for long prompts, 2K native resolution, and unified generation/editing [86]. Arena reports the model is in the Image Arena for comparison [87].

Industry Moves

Why it matters: Revenue run rates, enterprise spend share, and org stability (especially in open models) increasingly determine which models and tools become "defaults" in practice.

Anthropic: growth claims + enterprise spend signals

- Dario Amodei said Anthropic's revenue run rate went from ~**$100M** two years ago to **$19B** now [88].
- A separate post claims Anthropic is nearing a **$20B annual revenue run rate**, more than doubling from **$9B** at the end of 2025, and cites a valuation "around **$380B**" [89][90].
- Ramp-data commentary claims Anthropic commands the majority of U.S. business API spend and **>50% of enterprise AI subscription spend** (as of January), while OpenAI leads in business count [91][92][93].

[79] post by @perplexity_ai
[80] post by @perplexity_ai
[81] post by @AravSrinivas
[82] post by @Kling_ai
[83] post by @Kling_ai
[84] post by @Kling_ai
[85] post by @laszlogaal_
[86] post by @Alibaba_Qwen
[87] post by @arena
[88] post by @apoorv03
[89] post by @kimmonismus
[90] post by @kimmonismus
[91] post by @arakharazian
[92] post by @arakharazian
[93] post by @arakharazian

Alibaba Qwen: leadership departures + compute tension + market reaction

- One report says Alibaba CEO Eddie Wu held an emergency all-hands with the Qwen team, saying "I should have known about this sooner," amid tensions on restructuring, compute allocation, and model strategy [94][95]. It also cites an internal irony: external customers reportedly get smoother compute access than the internal team building Alibaba's "most important model" [96].
- A later post says Alibaba stock dropped **13.4%** this week and continued falling after key Qwen leaders announced departures, with doubts about Qwen 4 remaining a frontier open-source model without them [97][98].
- Separate commentary claims Qwen was used in **41%** of **7,692** AI papers on Hugging Face in 2025–2026, and at least **30%** monthly over a year (with May 2025 at 1 in 2 papers) [99].

Enterprise agent deployment: Factory + EY

- Factory says its partnership with EY will enable **10,000+ engineers** to ship software with autonomous agents, with adoption reportedly requiring throttling and repo restrictions [100][101].

Partnerships + funding

- **Spellbook** (legal AI) raised an additional **$40M** (on top of $50M raised last October), and reports serving **4,000+** legal teams/law firms in **80 countries** with **410 demos** booked in one week [102][103][104].
- **Cohere x Aston Martin F1:** Cohere announced a multi-year partnership; every team member gets access to Cohere's enterprise models and agentic AI platform [105][106].

Government direction-setting

- China's Premier Li Qiang outlined a 2026 AI agenda including "AI+" across industries, accelerating AI agent adoption, building ultra-large com-

[94] post by @poezhao0605
[95] post by @poezhao0605
[96] post by @poezhao0605
[97] post by @Yuchenj_UW
[98] post by @Yuchenj_UW
[99] post by @dongxi_nlp
[100] post by @FactoryAI
[101] post by @matanSF
[102] post by @scottastevenson
[103] post by @scottastevenson
[104] post by @scottastevenson
[105] post by @cohere
[106] post by @cohere

pute clusters, supporting public cloud, and promoting AI open-source communities [107].

Policy & Regulation

Why it matters: Legal and governance frameworks are increasingly binding constraints on what can be deployed (and where), especially for generative outputs and government use.

- **Frontier AI safeguards:** Yoshua Bengio endorsed the Human Statement and warned that frontier AI development is accelerating faster than safeguards, posing risks to democracy and society [108][109].
- **Copyright and AI authorship:** A cited episode summarizes the ongoing AI/copyright debate and references **Thaler v. Perlmutter** as a key case in the U.S. Copyright Office's refusal to register AI-generated works, with status noted for U.S. Supreme Court docket **25-449** (as of Feb. 25, 2026) [110][111].
- **Science funding leadership:** @regardthefrost (Jim) said he was nominated by President Trump to serve as Director of the National Science Foundation, calling for rigorous, replicable science and for government to take bigger financial risks on deeper questions [112].

Quick Takes

Why it matters: Smaller signals often foreshadow where teams will invest next.

- OpenAI released a new repo called **Symphony**: https://github.com/openai/symphony [113].
- OpenAI also released a repo called **Agent Plugins** (no details in the shared post) [114].
- **BullshitBench v2** tests nonsense detection; only Claude and Qwen 3.5 are said to score meaningfully above 60%, and "think harder" reasoning variants reportedly do worse by rationalizing nonsense [115][116].
- **SWE-bench** reached **1M weekly downloads**, with a "big update" coming to make it easier to run and support new benchmarks built on top [117].
- **OpenAI deprecates SWE-Bench Verified** due to contamination and flawed remaining tests, per a cited summary [118].

[107] post by @poezhao0605
[108] post by @Yoshua_Bengio
[109] post by @Yoshua_Bengio
[110] post by @LearnOpenCV
[111] post by @LearnOpenCV
[112] post by @regardthefrost
[113] post by @scaling01
[114] post by @scaling01
[115] post by @kimmonismus
[116] post by @kimmonismus
[117] post by @OfirPress
[118] post by @latentspacepod

- **Together AI / SSD** joins a broader inference-efficiency conversation: one researcher predicts inference compute will exceed training by decade's end, with premiums paid for lower latency and no one-size-fits-all stack across cloud vs edge [119][120][121].
- **Qdrant** joined Google's Agent Development Kit integrations ecosystem for persistent semantic memory and vector search in agent workflows [122].

Sources

1. post by @steph_palazzolo
2. post by @kimmonismus
3. post by @Techmeme
4. post by @legit_api
5. post by @kimmonismus
6. post by @kimmonismus
7. post by @BoWang87
8. post by @slow_developer
9. post by @washingtonpost
10. post by @mmitchell_ai
11. post by @steph_palazzolo
12. post by @ns123abc
13. post by @DeItaone
14. post by @spectatorindex
15. post by @TheRundownAI
16. post by @OpenAIDevs
17. post by @OpenAIDevs
18. post by @reach_vb
19. post by @scaling01
20. post by @kevinkern
21. post by @FactoryAI
22. post by @matanSF
23. post by @AmnonShashua
24. post by @bfl_ml
25. post by @robrombach
26. post by @GoogleResearch
27. post by @_arohan_
28. post by @tanishqkumar07
29. post by @tri_dao
30. post by @dair_ai
31. post by @omarsar0

[119] post by @awnihannun
[120] post by @awnihannun
[121] post by @awnihannun
[122] post by @qdrant_engine

32. post by @arcinstitute
33. post by @GoodfireAI
34. post by @BoWang87
35. post by @vicapow
36. post by @kevinweil
37. post by @code
38. post by @pierceboggan
39. post by @cursor_ai
40. post by @elicitorg
41. post by @elicitorg
42. post by @NotebookLM
43. post by @Google
44. post by @perplexity_ai
45. post by @AravSrinivas
46. post by @Kling_ai
47. post by @laszlogaal_
48. post by @Alibaba_Qwen
49. post by @arena
50. post by @apoorv03
51. post by @kimmonismus
52. post by @arakharazian
53. post by @poezhao0605
54. post by @Yuchenj_UW
55. post by @dongxi_nlp
56. post by @scottastevenson
57. post by @cohere
58. post by @poezhao0605
59. post by @Yoshua_Bengio
60. post by @LearnOpenCV
61. post by @regardthefrost
62. post by @scaling01
63. post by @kimmonismus
64. post by @OfirPress
65. post by @latentspacepod
66. post by @awnihannun
67. post by @qdrant_engine