GPT-5.4 benchmarking sharpens, security agents accelerate, and agent tooling expands

AI High Signal Digest

2026-03-07

GPT-5.4 benchmarking sharpens, security agents accelerate, and agent tooling expands

By AI High Signal Digest • March 7, 2026

Benchmarking and rollout details for GPT-5.4/Pro (including costs, long-context, and reliability trade-offs), major security developments (Firefox vulnerability research and Codex Security), plus a sweep of new agent tooling, open models, and infrastructure moves shaping deployment.

Top Stories

1) GPT-5.4's benchmark profile: bigger context, broad gains—and a higher bill

Why it matters: The latest third-party evaluations suggest GPT-5.4 is meaningfully stronger across science/coding/tool use/long-context tasks, but the cost curve (and some reliability metrics) moved in the wrong direction.

- **Artificial Analysis Intelligence Index:** GPT-5.4 (xhigh) ties for #1 at **57**, matching **Gemini 3.1 Pro Preview** and up from GPT-5.2 (xhigh) at 51 [1].
- **Context window + reasoning modes:** GPT-5.4 is reported with a **1.05M token** context window (up from 400K in GPT-5.2) and five reasoning effort modes (none → xhigh) [2].
- **Broad benchmark gains (with one notable regression):** Improvements vs GPT-5.2 (xhigh) include CritPt (+8 p.p.), TerminalBench Hard (+11 p.p.), HLE (+6 p.p.), ^{2}-Bench (+7 p.p.), SciCode (+5 p.p.), GPQA (+2 p.p.), and LCR (+1 p.p.); the only regression noted is **IFBench (-2 p.p.)** [3].

[1] post by @ArtificialAnlys
[2] post by @ArtificialAnlys
[3] post by @ArtificialAnlys

- **Cost / efficiency trade-off:** Despite modest token efficiency gains vs GPT-5.2, Artificial Analysis estimates the cost to run its full Intelligence Index rises ~**28%** to ~**$2,951** for GPT-5.4, and is ~3× Gemini 3.1 Pro Preview (~$892), driven by both token usage and higher per-token prices [4][5][6][7].
- **Accuracy vs hallucinations tension (AA-Omniscience):** GPT-5.4 improves accuracy (44% → 50%) but shows a worse hallucination rate (80% → **89%**) attributed to a higher attempt rate (91% → 97%) [8].

Full model card/results: https://artificialanalysis.ai/models/gpt-5-4 [9]

2) GPT-5.4 Pro hits a new SOTA on CritPt—at a steep "reasoning premium"

Why it matters: CritPt is positioned as research-level physics reasoning with a private dataset; the jump to 30% in ~4 months is notable, but it also highlights a widening gap between **best-possible** results and **economically deployable** results.

- Artificial Analysis reports **GPT-5.4 Pro (xhigh)** reaching **30%** on CritPt, a **10-point gain** over the prior best of 9% when CritPt launched in Nov 2025 [10].
- The same evaluation is described as costing **over $1k**, about **13×** GPT-5.4 (xhigh), driven by output pricing (**$180/1M output tokens** vs $15) despite similar token counts (6.0M vs 5.5M) [11].
- Separate commentary flags the cost delta: GPT-5.4-Pro-xhigh is reported as **13.275×** more expensive than GPT-5.4-xhigh [12].

CritPt leaderboard link: https://artificialanalysis.ai/evaluations/critpt?models=gpt-oss-120b%2Cgpt-5-4-pro%2Cgpt-5-2%2Cgpt-5-3-codex%2Cgpt-5-4%2Cgemini-3-1-pro-preview%2Cgemini-3-flash-reasoning%2Cclaude-4-5-haiku-reasoning%2Cclaude-opus-4-6-adaptive%2Cclaude-sonnet-4-6-adaptive%2Cmistral-large-3%2Cdeepseek-v3-2-reasoning%2Cdeepseek-v3-2-speciale%2Cgrok-4-1-fast-reasoning%2Cminimax-m2-5%2Ckimi-k2-5%2Cglm-5%2Cqwen3-5-397b-a17b%2Cgemini-3-pro [13]

[4] post by @ArtificialAnlys
[5] post by @ArtificialAnlys
[6] post by @ArtificialAnlys
[7] post by @ArtificialAnlys
[8] post by @ArtificialAnlys
[9] post by @ArtificialAnlys
[10] post by @ArtificialAnlys
[11] post by @ArtificialAnlys
[12] post by @scaling01
[13] post by @ArtificialAnlys

3) "Security agents" are becoming a headline capability: Firefox vulnerability research + Codex Security

Why it matters: The same frontier-model capabilities improving coding and tool use are translating into vulnerability discovery at scale—raising the bar for defense (and shrinking the window before exploitation improves).

- **Claude Opus 4.6 on Firefox (Anthropic × Mozilla):** Anthropic says it partnered with Mozilla to test Claude's ability to find vulnerabilities in Firefox, reporting **22 vulnerabilities** found in two weeks, including **14 high-severity** (about **one-fifth** of Mozilla's 2025 high-severity remediations) [14].
- Anthropic also warns that while models are "currently better at finding vulnerabilities than exploiting them," the gap is "unlikely to last," urging developers to improve software security [15].
- A separate summary reports that in exploitation testing, Claude produced a working browser exploit **twice** (after several hundred attempts and about **$4,000** in API credits) on a stripped test system, and frames vulnerability finding as ~**10× cheaper** than exploiting "for now" [16][17].

In parallel, **OpenAI introduced Codex Security**, an application security agent that finds vulnerabilities, validates them, and proposes fixes for review and patching [18]. OpenAI says it evolved from **Aardvark** (private beta last year) and improved signal quality (reduced noise/false positives, better severity accuracy) [19][20][21].

4) LisanBench "Thinking" results surge; benchmark creator considers making it harder

Why it matters: These results are another datapoint that *reasoning-budgeted* variants can dominate certain open-ended tasks—while also showing how quickly some benchmarks can saturate.

- Latest LisanBench "Thinking (16k)" top scores include **Opus 4.6 Thinking (14083)** and **Sonnet 4.6 Thinking (11789.67)**, followed by **Gemini 3.1 Pro (high) 6414.67**; **GPT-5.4 (medium)** is listed at **5273.33** [22].
- The benchmark creator says they may "either make a harder version of LisanBench or discontinue it" [23], and separately notes that with

[14] post by @AnthropicAI
[15] post by @AnthropicAI
[16] post by @TheRundownAI
[17] post by @TheRundownAI
[18] post by @OpenAIDevs
[19] post by @OpenAIDevs
[20] post by @OpenAI
[21] post by @OpenAIDevs
[22] post by @scaling01
[23] post by @scaling01

Opus/Sonnet 4.6 it "seems like it's saturating," leaving "only reasoning efficiency" measurable beyond a point [24][25].

5) Compute spending and infrastructure expansion continues to accelerate

Why it matters: The capex and physical buildout signal how aggressively the industry is committing to scaling—even as model lifecycles stay short and evaluation costs rise.

- One estimate claims **MSFT, AMZN, META, GOOG** will spend **$650B** this year [26].
- A separate roundup flags **SoftBank seeking up to $40B** in a loan mostly to finance its OpenAI stake [27].
- OpenAI infrastructure: construction is underway at a **Port Washington, Wisconsin** site with **VantageDC** and **Oracle**, described as part of OpenAI's long-term compute strategy; the "first steel beams went up" this week [28][29].

Research & Innovation

Why it matters: This cycle's research points to three themes: (1) better **efficiency** (architectures/training), (2) more **agent-realistic evaluation**, and (3) new approaches to **memory and continual learning**.

Hybrid architectures and data efficiency

- **Allen AI:** Reports a key finding that **hybrid models** can be "substantially more data-efficient than transformers," with **Olmo Hybrid** matching **Olmo 3** on MMLU using **49% fewer tokens** (~2× efficiency) [30][31].
- Lambda published a model card with speed tests for `olmo-hybrid-instruct-dpo-7b` across A100/H100/B200 [32][33].

Compact multimodal reasoning for practical agents

- **Microsoft Phi-4-reasoning-vision-15B:** A 15B parameter multimodal reasoning model combining visual understanding with

[24] post by @scaling01
[25] post by @scaling01
[26] post by @scaling01
[27] post by @TheStalwart
[28] post by @sk7037
[29] post by @sk7037
[30] post by @allen_ai
[31] post by @allen_ai
[32] post by @TheZachMueller
[33] post by @TheZachMueller

structured reasoning over text and images, aimed at the capability/efficiency "sweet spot" for practical agent deployments [34][35][36]. Paper: https://arxiv.org/abs/2603.03975 [37].

Benchmarks for more realistic "software engineering" agents

- **SWE-CI:** A new benchmark designed around continuous integration workflows (running test suites, catching regressions, maintaining code quality across multiple changes), positioned as a step beyond single-issue bug-fix benchmarks [38][39]. Paper: https://arxiv.org/abs/2603.03823 [40].

Continual learning + instant specialization via LoRA hypernetworks

- **Sakana AI Labs:** Introduced **Doc-to-LoRA** (turning documents into memory) and **Text-to-LoRA** (turning task descriptions into behavior adapters) using a hypernetwork that generates LoRA weights; meta-training takes days/weeks, but adapter generation is milliseconds at runtime [41][42]. Claimed benefits include long-term memory without re-reading documents and "instant task specialization" without a fine-tuning pipeline [43][44].

Fine-tuning efficiency and "forgotten knowledge"

- A research note claims **replaying generic pre-training data** during fine-tuning improves data efficiency, reduces forgetting, and can improve performance on the fine-tuning domain (especially when that domain is scarce in pre-training) [45][46].
- Separate work notes that a drop in prior-task performance in VLAs doesn't necessarily mean knowledge is gone; it can be "rapidly recovered with minimal finetuning" [47].

Language and speech data availability

- **Google Research WAXAL:** Open-access dataset with **2,400+ hours** of speech data for **27 Sub-Saharan African languages** serving **100M+**

[34] post by @omarsar0
[35] post by @omarsar0
[36] post by @omarsar0
[37] post by @omarsar0
[38] post by @dair_ai
[39] post by @dair_ai
[40] post by @dair_ai
[41] post by @TheTuringPost
[42] post by @TheTuringPost
[43] post by @TheTuringPost
[44] post by @TheTuringPost
[45] post by @kothasuhas
[46] post by @percyliang
[47] post by @huihan_liu

speakers, positioned as addressing data scarcity across Africa's **2000+ spoken languages** [48][49]. Dataset: http://goo.gle/4cxNHae [50].

Products & Launches

Why it matters: Agent tooling is expanding along three fronts: (1) security and code maintenance, (2) "computer" orchestration and automation, and (3) creative workflows that are composable and model-agnostic.

Security + open source maintenance

- **Codex Security (research preview):** OpenAI's application security agent is in research preview [51]. OpenAI says it's rolling out to ChatGPT **Enterprise/Business/Edu** via Codex web with **free usage for the next month** [52], and is now also available on **ChatGPT Pro** accounts [53].
- **Codex for Open Source:** OpenAI is launching Codex for OSS maintainers to help with code review, understanding large codebases, and strengthening security coverage [54]. Maintainers receive **API credits**, **6 months of ChatGPT Pro with Codex**, and **access to Codex Security** as needed [55]. Apply: http://developers.openai.com/codex/community/codex-for-oss [56].

Agent "computer" platforms add reuse and automation

- **Perplexity Computer:** Shipped Voice Mode, Skills, Model Council, and added GPT-5.4 / GPT-5.4 Thinking (including as an orchestrator model) [57][58]. Perplexity also demoed generating a formatted Excel spreadsheet with live macro indicators from a simple prompt plus a Federal Reserve API key [59].
- **Claude Code desktop:** Launched **local scheduled tasks**, letting users run regular tasks while the computer is awake [60].

[48] post by @GoogleResearch
[49] post by @GoogleResearch
[50] post by @GoogleResearch
[51] post by @OpenAI
[52] post by @OpenAIDevs
[53] post by @OpenAIDevs
[54] post by @OpenAIDevs
[55] post by @kevinweil
[56] post by @OpenAIDevs
[57] post by @AravSrinivas
[58] post by @AravSrinivas
[59] post by @AskPerplexity
[60] post by @trq212

Creative + multimodal workflows

- **NotebookLM:** Google says it can turn sources into "cinematic video explainers," with Cinematic Video Overviews rolling out for Ultra users in English [61][62].
- **Hugging Face Modular Diffusers:** New Diffusers submodule enabling composable diffusion pipelines (mix-and-match blocks; visual workflow via Mellon; share custom blocks on HF Hub), with a commitment to maintain both the classic `DiffusionPipeline` and new `ModularPipeline` abstractions [63][64][65][66][67]. Blog: https://huggingface.co/blog/modular-diffusers [68].

Developer-facing tools and marketplaces

- **T3 Code:** A fully open-source tool built on Codex CLI, intended to scale parallel agent workflows beyond what CLIs handle well; available at http://t3.codes or via `npx t3@alpha` [69][70][71].
- **Anthropic Claude marketplace:** Anthropic says organizations can apply existing spend commitments toward Claude-powered partner solutions (e.g., GitLab, Harvey, Replit, Snowflake) [72].

Industry Moves

Why it matters: Distribution (where models show up), pricing/subsidies, and infrastructure decisions are increasingly shaping adoption as much as raw benchmark performance.

"Coding model arms race" intensifies

- **Cursor:** Reported mandate labeled "P0 #1" to "Build the best coding model" [73].
- **Claude Code subsidization (as inferred from Cursor analysis):** A $200/month plan reportedly moved from allowing ~$2,000 of compute to ~$5,000 (2.5×) [74].

[61] post by @Google
[62] post by @Google
[63] post by @RisingSayak
[64] post by @RisingSayak
[65] post by @RisingSayak
[66] post by @RisingSayak
[67] post by @RisingSayak
[68] post by @RisingSayak
[69] post by @theo
[70] post by @theo
[71] post by @theo
[72] post by @claudeai
[73] post by @tanayj
[74] post by @bearlyai

Open models and regional ecosystems

- **Sarvam AI:** Open-sourced two India-built reasoning models (**Sarvam 30B** and **105B**) with an emphasis on full-stack in-house work (data, training, RL, tokenizer design, inference optimization) and performance in Indian languages; weights are available on Hugging Face and AIKosh, with SGLang day-0 support and vLLM support "coming soon" [75][76].

Developer tooling + enterprise deployments

- **ToyotaGPT:** Toyota Motor North America equipped **56,000 employees** with ToyotaGPT built on LangGraph [77].
- **Databricks:** Announced day-one access to GPT-5.4 on Databricks [78].

Geographic clustering

- A London-focused roundup claims OpenAI plans London as its largest research hub outside San Francisco, while Anthropic, xAI, Microsoft, DeepMind, Perplexity, Groq, and Cursor are also expanding or establishing major presence there [79].

Policy & Regulation

Why it matters: Government procurement decisions and legal challenges are becoming first-order constraints on which models can be used (and where), especially in defense contexts.

Anthropic vs. Department of War: "supply chain risk" designation and fallout

- Anthropic says the Department of War's supply-chain risk designation is narrower than early headlines suggested, affecting only Claude's direct use in certain Department-linked contracts, while most customers remain unaffected [80]. Anthropic CEO Dario Amodei calls the move legally shaky, says Anthropic will fight it in court, and reiterates support for U.S. national security—offering models at nominal cost during a transition to avoid disrupting critical operations [81].
- Separately, Emil Michael states there is **"no active Department of War negotiation with Anthropic"** [82].

[75] post by @pratykumar
[76] post by @pratykumar
[77] post by @LangChain
[78] post by @databricks
[79] post by @thealexbanks
[80] post by @kimmonismus
[81] post by @kimmonismus
[82] post by @USWREMichael

- Google is reported as saying Anthropic will remain available for **non-defense workloads** on Google Cloud [83].

Privacy litigation signal

- A roundup flags Meta's AI glasses being hit with a privacy suit (details linked) [84][85].

Quick Takes

Why it matters: These are smaller datapoints that still shift day-to-day practice (what wins on real tasks, what breaks, and what teams deploy next).

- **TaxCalcBench:** GPT-5.4 scores **56.86%** perfect tax returns, #1 overall and above Claude Opus 4.6 (52.94%); a separate post cites a jump from GPT-5.2 (34%) to GPT-5.4 (57%) [86][87].
- **LiveBench:** GPT-5.4-xhigh takes 1st place with very strong reasoning and coding scores [88].
- **Arena (text):** GPT-5.4 High lands in the top 10 Text Arena, described as substantially more rounded than GPT-5.2 High with large gains in categories like creative writing and legal/government [89][90][91].
- **Kaggle challenges:** A claim that GPT-5.4 is almost 2× as good as GPT-5.2 at Kaggle challenges requiring designing/building/training ML models on GPUs (success = bronze medal or better) [92].
- **"Tiny program" demo:** GPT-5.4 reportedly generates a <5000-byte C program to run GPT-2 inference from raw weights in under 15 minutes [93].
- **Prompt-injection incident:** An attacker reportedly stole an npm token by injecting a prompt into a GitHub issue title that an AI triage bot executed [94].
- **Model execution speed:** Mercury 2 (diffusion, not autoregressive) claims **1,009 tokens/sec**, targeting agent workflows where latency stacks up [95].

[83] post by @PolymarketMoney
[84] post by @TheRundownAI
[85] post by @TheRundownAI
[86] post by @michaelrbock
[87] post by @scaling01
[88] post by @scaling01
[89] post by @arena
[90] post by @arena
[91] post by @arena
[92] post by @_simonsmith
[93] post by @markchen90
[94] post by @zats
[95] post by @yupp_ai

- **vLLM attention portability:** vLLM's Triton attention backend (~800 lines) is presented as cross-platform across NVIDIA/AMD/Intel; it matches SOTA on H100 and is ~5.8× faster than earlier implementations on MI300, and is now the default on AMD ROCm [9697].

Sources

1. post by @ArtificialAnlys
2. post by @ArtificialAnlys
3. post by @ArtificialAnlys
4. post by @ArtificialAnlys
5. post by @ArtificialAnlys
6. post by @ArtificialAnlys
7. post by @scaling01
8. post by @ArtificialAnlys
9. post by @AnthropicAI
10. post by @AnthropicAI
11. post by @TheRundownAI
12. post by @OpenAIDevs
13. post by @OpenAIDevs
14. post by @OpenAI
15. post by @scaling01
16. post by @scaling01
17. post by @scaling01
18. post by @scaling01
19. post by @TheStalwart
20. post by @sk7037
21. post by @sk7037
22. post by @allen_ai
23. post by @TheZachMueller
24. post by @TheZachMueller
25. post by @omarsar0
26. post by @dair_ai
27. post by @TheTuringPost
28. post by @kothasuhas
29. post by @percyliang
30. post by @huihan_liu
31. post by @GoogleResearch
32. post by @GoogleResearch
33. post by @OpenAI
34. post by @OpenAIDevs
35. post by @OpenAIDevs

[96] post by @vllm_project
[97] post by @vllm_project

36. post by @OpenAIDevs
37. post by @kevinweil
38. post by @AravSrinivas
39. post by @AskPerplexity
40. post by @trq212
41. post by @Google
42. post by @Google
43. post by @RisingSayak
44. post by @RisingSayak
45. post by @RisingSayak
46. post by @RisingSayak
47. post by @RisingSayak
48. post by @theo
49. post by @theo
50. post by @theo
51. post by @claudeai
52. post by @tanayj
53. post by @bearlyai
54. post by @pratykumar
55. post by @LangChain
56. post by @databricks
57. post by @thealexbanks
58. post by @kimmonismus
59. post by @USWREMichael
60. post by @PolymarketMoney
61. post by @TheRundownAI
62. post by @TheRundownAI
63. post by @michaelrbock
64. post by @scaling01
65. post by @scaling01
66. post by @arena
67. post by @_simonsmith
68. post by @markchen90
69. post by @zats
70. post by @yupp_ai
71. post by @vllm_project