# OpenAI cost-and-growth projections, Claude Code Security, and new signals on long-horizon agents

AI High Signal Digest

2026-02-21

## OpenAI cost-and-growth projections, Claude Code Security, and new signals on long-horizon agents

*By AI High Signal Digest • February 21, 2026*

Key developments include circulated projections of OpenAI's sharply higher revenue targets alongside massive projected compute/training spend, Anthropic's Claude Code Security launch (and immediate market reaction), and new METR time-horizon estimates—plus fresh signals in specialized inference hardware and local AI tooling consolidation.

### Top Stories

**1) OpenAI's updated financial outlook: higher revenue targets, higher compute spend, and margin pressure**

*Why it matters:* The numbers being discussed imply a strategy that depends on **massive infrastructure buildout** and continued product growth—while acknowledging near-term margin constraints.

A set of circulated projections says OpenAI raised its **5-year revenue forecast by ~27%**, with **2025 revenue at $13.1B** (tripled YoY; +$100M vs projections) and new targets of **$30B (2026)** and **$62B (2027)** [1]. The same summary claims consumer sales could reach **$17B in 2026** and **$150B by 2030** (more than half of total revenue) [1].

On costs, the projections highlight sharply rising spend:

- Total spend: **$25B (2026)** and **$57B (2027)** [1]
- Infrastructure/compute: **$665B through 2030** [1]

- Training: **$8.3B (2025) → $32B (2026) → $65B (2027)**; nearly **$440B through 2030** [1]
- Inference: **$8B (2025) → $14B (2026) → $26B (2027)** [1]

The same thread attributes a 2025 adjusted gross margin miss to expensive last-minute compute purchases, noting adjusted gross margin fell from **40% to 33%** against a **46%** target [1]. It also claims OpenAI ended 2025 with ~**$40B cash** and expects to be **cash-flow positive by 2030** [1].

Separately, another post summarizing The Information reporting says OpenAI ended 2025 with **$13.1B revenue** and **$8B cash burn**, targeting **$284B revenue in 2030** alongside **$665B computing costs through 2030**, and notes OpenAI is trying to raise **>$100B** [2]. (A related post also cites "cash burn through 2030" now **2×** prior estimates, totaling **$112B**.) [3, 4]

Growth metrics in the projections include a reported new peak of **910M weekly active users** for ChatGPT, with growth said to have slowed in fall 2025 but re-accelerated after **GPT-5.1/5.2** updates; the long-term goal cited is **2.75B WAUs by 2030** [1].

### 2) Anthropic ships Claude Code Security (research preview); market reactions highlight "AI scans first" dynamics

*Why it matters:* AI-assisted vulnerability discovery and patch suggestions could compress security cycles—and early reactions suggest investors are taking "AI security tooling" seriously.

Anthropic introduced **Claude Code Security** in limited research preview [5]. It scans codebases for vulnerabilities and suggests **targeted patches for human review**, aiming to catch issues that traditional tools miss [5, 6]. Anthropic says the system (powered by **Claude Opus 4.6**) found **500+ vulnerabilities** in production open-source codebases, including bugs "hidden for decades" [7, 6].

Anthropic positions the product as reasoning over code "like a human security researcher," tracing data flows and component interactions (not just pattern matching), and re-checking its own findings to reduce false positives—while requiring human approval before anything is applied [6].

Links and access:

- Product details: https://www.anthropic.com/news/claude-code-security [5]
- Waitlist: https://claude.com/contact-sales/security [7]
- Posts also note open-source maintainers are encouraged to apply for free, expedited access [8].

Market reaction was widely noted: one post claims cybersecurity stocks dropped within 30 minutes of the announcement (e.g., **CrowdStrike -6%**, **Cloudflare -5.2%**, **Palo Alto -3.8%**, **Zscaler -3%**, **Fortinet -3%**) [9]. Another post describes **$10B** in market cap loss within an hour, listing **CrowdStrike -6.5%**, **Cloudflare -6%**, **Okta -5.7%** [10].

### 3) METR time-horizon updates: Opus 4.6 leads on point estimate; GPT-5.3-Codex measured; multiple cautions on interpretation

*Why it matters:* "Longer time horizons" are one of the clearest operational signals for agents doing multi-step work—but the benchmark's uncertainty and saturation mean the headline numbers can be misleading.

METR estimates **Claude Opus 4.6** has a **50% time horizon of ~14.5 hours** on software tasks (95% CI: **6–98 hours**), noting the measurement is "extremely noisy" because the current suite is nearly saturated [11]. METR also estimates **GPT-5.3-Codex** (`high` reasoning effort) at ~**6.5 hours** (95% CI: **3–17 hours**) and says OpenAI provided API access for the evaluation [12].

METR notes it used its Triframe scaffold (not Codex), and a partial run with a Codex scaffold looked similar—consistent with past comparisons [13]. It adds that initial scaffolding/format issues hurt performance, and after addressing them, it had the impression this model may be **more scaffold-sensitive** than usual [14].

Several researchers and practitioners urged caution:

- One commenter says slight task distribution changes could have yielded **8 hours or 20 hours** for Opus 4.6, underscoring how noisy the estimate is [15].
- Another notes METR's newest points are "weaker updates" than earlier ones due to saturation/limited long-duration tasks, while still supporting the view that progress hasn't slowed significantly [16, 17].
- A separate explanation highlights why small reductions in near-zero per-step error rates can look like a large "time horizon" jump (e.g., in a **1000-step** task, **1%** per-step error implies ~**37%** success, while **0.5%** implies ~**61%**) [18].

### 4) Taalas launches model-specialized silicon for ultra-fast inference—alongside skepticism about scaling

*Why it matters:* If "model-specific" chips deliver large cost/latency gains, they could materially change inference economics for certain workloads; the hard question is whether the approach scales to larger models and long contexts.

Taalas announced its first product after **$30M** in development by **24 people**, emphasizing specialization, speed, and power efficiency [19]. Multiple posts report Taalas running **Llama-3.1-8B** at roughly **17k tokens/sec** (or ~16k tokens/sec) per user with low latency [20, 21]. One post summarizes the key idea as: each chip is specialized to a given model—"**the chip is the model**" [21].

Access points were shared:

- Details: https://taalas.com/the-path-to-ubiquitous-ai/ [19]
- Demo chatbot: https://chatjimmy.ai [19]
- API request: https://taalas.com/api-request-form/ [19]

Economics and constraints were debated. An analysis argues spending "tens of millions" on tape-out can make financial sense if it yields ~**10× efficiency** and ~**1/10th latency**, but flags tape-out latency (e.g., "**2 months is too slow**" vs rapid model iteration) and suggests a hybrid approach (base model in silicon, adapters post-trained) [22]. Others called out scaling limits to "big models" and "large contexts" [23], and separately described 8B/short-context speed as a "parlor trick" unless it extends to very long contexts [24].

### 5) Hugging Face + ggml/llama.cpp join forces to accelerate "local AI"

*Why it matters:* This is a distribution + tooling alignment around running models on users' own hardware, which some frame as a counterweight to cloud-driven centralization.

ggml.ai (the team behind **llama.cpp**) announced it is joining Hugging Face [25]. The stated joint mission is to continue building **ggml**, make **llama.cpp** more accessible, and "make local AI easy and efficient to use by everyone on their own hardware" [25].

## Research & Innovation

*Why it matters:* This cycle's research emphasizes **agents (coordination, autonomy, long-horizon training)** and **systems/attention** improvements that make models more useful under real constraints.

- **DreamDojo (open-source robot world model):** NVIDIA's DreamDojo is described as an interactive robot world model pretrained on **44K hours** of human egocentric video, with a distilled real-time version running at **10 FPS** and enabling live teleop, policy evaluation, and model-based planning; one reported result is **+17% real-world success** on a fruit packing task out of the box [26]. Project/paper/code links were shared: https://dreamdojo-world.github.io/ and https://arxiv.org/abs/2602.06949 and https://github.com/NVIDIA/DreamDojo [27].
- **Long-horizon agent training (KLong):** KLong proposes a two-phase method—trajectory-splitting supervised fine-tuning followed by progressive RL with staged timeouts—to address context loss and sparse rewards over long trajectories [28]. A "Research-Factory" pipeline is described as generating thousands of long-horizon training trajectories from Claude 4.5 Sonnet [28].

- **Long-context retrieval/reasoning (LUCID):** LUCID is presented as a new attention mechanism to improve retrieval and reasoning in long-context LLMs [29].
- **Multi-agent orchestration selection:** A paper introduces task-adaptive orchestration that chooses among four canonical agent topologies (parallel, sequential, hierarchical, hybrid) based on task dependency graphs, reporting **12–23%** improvements over static single-topology baselines [30].
- **Collective behavior at scale (social dilemmas):** New research proposes an evaluation framework for hundreds of LLM agents in social dilemmas; reported findings include that "newer, more capable models" can lead to worse societal outcomes as individually optimizing agents drive populations to poor equilibria, with larger populations amplifying the risk [31]. Paper: https://arxiv.org/abs/2602.16662 [31].
- **Dynamic populations in RL (Fluid-Agent RL):** Fluid-Agent RL allows agents to dynamically create additional agents during an episode (non-fixed populations), with game-theoretic solution concepts validated on fluid variants of Predator-Prey and Level-Based Foraging [32].
- **Agent training data standardization (ADP):** Agent Data Protocol (ADP) was accepted as an **ICLR 2026 Oral** and expanded to **3.2M instances**, with support for **3M trajectories** and added datasets (SWE-Play, MiniCoder, Toucan) [33, 34].
- **Video diffusion architecture insight:** Research claims causality is separable from denoising in causal video diffusers, with lower layers handling noise-level-independent causal processing and upper layers focusing on intra-frame denoising; separating them is said to bring practical and speed benefits [35, 36].

## Products & Launches

*Why it matters:* The biggest user impact comes from capabilities packaged into workflows—especially for coding, security, research, and media generation.

- **Claude Code desktop updates:** Claude Code on desktop can now preview running apps, review code, and handle CI failures and PRs in the background [37].
- **Anthropic CLI signals:** A new GitHub repo for **anthropic-cli** surfaced, and an Anthropic employee said they're working on launching a CLI for the Claude API [38, 39].
- **Perplexity weekly ship log:** Perplexity listed updates including Comet pre-ordering on iOS, support for Claude Sonnet 4.6 and Gemini 3.1 Pro, response preferences, Enterprise Memory, a personalized Comet Assistant, and auditable financials with SEC links [40].

- **Runway model hub:** Runway says "all of the world's best models" are available inside Runway, listing Kling 3.0/2.6/2.5 variants, WAN2.2 Animate, GPT-Image-1.5, and Sora 2 Pro (with more coming) [41].
- **Pika "AI Selves":** Pika introduced "Pika AI Selves," described as customizable AI beings users "birth, raise, and set loose" as living extensions with persistent memory [42]. Waitlist: pika.me [42].
- **Aristotle (AI co-scientist):** "Aristotle" launched as a next-generation AI co-scientist, now live and free for verified U.S. researchers, with models including X1 Verify, X1 Search, and X1 Spark [43, 44].
- **LlamaIndex workflow builders:** LlamaIndex released LlamaAgent Builder (describe document-agent workflows in natural language) and highlighted LlamaExtract for agentic extraction (e.g., schema creation with bounding-box tracebacks to source text) [45, 46].
- **Artificial Analysis Image Lab:** Image Lab allows users to run a single prompt across up to **25 image models**, generating up to **20 images** per model and viewing results quickly; free trial link shared [47, 48].
- **Ollama 0.16.3:** Ollama shipped version 0.16.3 with **Cline** and **Pi** integrations out of the box (e.g., `ollama launch cline`) [49].

## Industry Moves

*Why it matters:* Device strategy, distribution, and operational mishaps often determine who captures demand—independent of benchmark performance.

- **OpenAI device roadmap (reported):** Posts summarizing reporting say OpenAI has a **200-person team** building a family of AI-powered devices, starting with a **$200–$300 smart speaker** designed with **Jony Ive's LoveFrom**, featuring a **camera**, environmental awareness, Face ID-style purchasing, and proactive "nudges," with release **no earlier than February 2027** [50, 51].
- **Codex growth in India:** Sam Altman said India is OpenAI's **fastest growing market for Codex**, with weekly users up **4× in the past two weeks**, and noted a meeting with PM Narendra Modi about energy around AI in India [52].
- **OpenAI developer community:** OpenAI is organizing **Codex meetups** via an ambassador community, with a central meetups page shared for cities and events [53].
- **Amazon internal coding assistant outage (reported):** A post citing an FT article says Amazon's internal AI coding assistant deleted existing code to start from scratch, causing part of AWS to go down for **13 hours**, and that it wasn't the first time [54].

## Policy & Regulation

*Why it matters:* AI deployment increasingly depends on auditability, compute access programs, and secure tool-use boundaries.

- **Independent AI auditing standards:** A nonprofit, the AI Verification and Research Institute (Averi), aims to establish standards for independent audits of AI systems to evaluate risks like misuse, data leaks, and harmful behavior [55].
- **Academic compute access (Google TPUs):** Google is launching the **2026 Google TPU Research & Education Awards**, offering free access to latest TPUs, an unrestricted funding gift for grad student support, and Google Cloud credits [56]. Apply: https://goo.gle/2026-tpu-rfp [56].
- **Tool-calling access-control vulnerability (reported):** Piotr Czapla described an issue where an LLM given a list of allowed tools may attempt to call a tool that wasn't provided, undermining access control; a related post claims it impacts major US providers except OpenAI [57].

## Quick Takes

*Why it matters:* Smaller releases and benchmark notes often preview what developers will feel next—speed, cost, and workflow reliability.

- **Gemini 3 Deep Think:** Google announced a major upgrade to Gemini 3 Deep Think, positioning it as a specialized reasoning mode for frontier science/math/engineering with early API access for researchers and enterprises [58].
- **Gemma (edge):** Demis Hassabis said a new Gemma open-source model "very powerful for edge devices" will be released soon [59].
- **Codex speed:** GPT-5.3-Codex-Spark was reported ~**30% faster**, serving at **>1200 tokens/sec** [60].
- **Model leaderboard notes (Arena):** Alibaba's **Qwen3.5-397B-A17B** was described as a top-3 open model in Text Arena and tied top-2 open in Vision Arena, with Code Arena scores "coming soon" [61, 62, 63].
- **MiniMax usage (OpenRouter):** One post says MiniMax was the first model to break **3 trillion tokens in a week** on OpenRouter rankings [64].
- **vLLM tuning tip:** A performance note suggests SGLang can be faster than vLLM on some models because vLLM may choose DeepGemm; recommended setting `VLLM_USE_DEEP_GEMM=0` [65].
- **"Pro Lite" signal in ChatGPT web app code:** A post claims the ChatGPT web app code mentions a new "ChatGPT Pro Lite" plan [66].

---

### Sources

1. X post by @stalkermustang
2. X post by @srimuppidi
3. X post by @steph_palazzolo
4. X post by @steph_palazzolo
5. X post by @claudeai
6. X post by @LiorOnAI
7. X post by @_catwu
8. X post by @sammcallister
9. X post by @nrehiew_
10. X post by @TheGeorgePu
11. X post by @METR_Evals
12. X post by @METR_Evals
13. X post by @METR_Evals
14. X post by @METR_Evals
15. X post by @idavidrein
16. X post by @deanwball
17. X post by @idavidrein
18. X post by @xlr8harder
19. X post by @taalas_inc
20. X post by @AymericRoucher
21. X post by @awnihannun
22. X post by @awnihannun
23. X post by @teortaxesTex
24. X post by @teortaxesTex
25. X post by @ggerganov
26. X post by @DrJimFan
27. X post by @ShenyuanGao
28. X post by @dair_ai
29. X post by @dvsaisurya
30. X post by @omarsar0
31. X post by @omarsar0
32. X post by @dair_ai
33. X post by @yueqi_song
34. X post by @gneubig
35. X post by @SimulatedAnneal
36. X post by @sedielem
37. X post by @claudeai
38. X post by @scaling01
39. X post by @katelyn_lesse
40. X post by @tylertate
41. X post by @runwayml
42. X post by @pika_labs
43. X post by @autopoiesislab
44. X post by @kimmonismus

45. X post by @tuanacelik
46. X post by @jerryjliu0
47. X post by @ArtificialAnlys
48. X post by @ArtificialAnlys
49. X post by @ollama
50. X post by @kimmonismus
51. X post by @kimmonismus
52. X post by @sama
53. X post by @OpenAIDevs
54. X post by @MikeIsaac
55. X post by @DeepLearningAI
56. X post by @RisingSayak
57. X post by @jeremyphoward
58. X post by @dl_weekly
59. X post by @Hangsiin
60. X post by @thsottiaux
61. X post by @arena
62. X post by @arena
63. X post by @arena
64. X post by @alexatallah
65. X post by @TheZachMueller
66. X post by @btibor91