Codex Pricing, Anthropic Routing, and the Shift Toward Gated Cyber AI

AI High Signal Digest

2026-04-10

Codex Pricing, Anthropic Routing, and the Shift Toward Gated Cyber AI

By AI High Signal Digest • April 10, 2026

OpenAI reshaped ChatGPT pricing around Codex, Anthropic turned multi-model orchestration into a platform feature, and new research sharpened both the promise and limits of agent systems. This brief also tracks open-model momentum, enterprise infrastructure moves, and the policy shift toward controlled release of cyber-capable AI.

Top Stories

Why it matters: Product strategy, model competition, and deployment controls are shifting at the same time. The result is a market where coding agents are being monetized, orchestration is becoming a first-class product, and the most sensitive models are staying gated.

OpenAI creates a new Codex-heavy price tier

OpenAI said it is updating ChatGPT Pro and Plus to support growing Codex use, introducing a new **$100/month Pro tier** with **5x more Codex usage** than Plus, access to the exclusive Pro model, and unlimited Instant and Thinking models [1]. Through May 31, subscribers get up to **10x** Plus usage on Codex, while Plus is being rebalanced toward more sessions across the week rather than longer single-day sessions; the existing **$200 Pro** plan remains the highest-usage option [1, 2, 3].

In a recent discussion, OpenAI Devs' @reach_vb said Codex had reached **3M weekly users** [4].

Impact: OpenAI is no longer treating coding assistance as just another chat feature. It is building explicit pricing and usage tiers around agentic software

work.

Anthropic turns multi-model routing into a product feature

Anthropic brought its **advisor strategy** to the Claude Platform: Opus acts as the advisor, while Sonnet or Haiku executes, with the goal of near Opus-level intelligence at a fraction of the cost [5]. In Anthropic's evals, Sonnet with an Opus advisor scored **2.7 percentage points higher** on SWE-bench Multilingual than Sonnet alone while costing **11.9% less** per task [6]. Anthropic's Alex Albert said this "phone a friend" pattern improves performance while cutting total cost by reducing wasted tokens on hard tasks [7].

Impact: Model quality is no longer the whole story. How models are combined is becoming a competitive product surface.

Qwen and Gemma show the open stack still has momentum

Alibaba launched **Qwen3.6-Plus**, described as a frontier agentic coding model that matches or beats Claude Opus 4.5 on SWE-bench and Terminal-Bench 2.0 [8]. At the same time, Google DeepMind said **Gemma 4** outperforms models **10x its size** without massive compute and crossed **10M downloads** in its first week, with the Gemma family above **500M** total downloads [9]. Unsloth also said Gemma-4-31B can be fine-tuned for free on Kaggle and fits in roughly **22GB VRAM** on two free Tesla T4 GPUs [10, 11].

Impact: Frontier pressure is not only coming from closed U.S. labs. Strong coding performance and easier access are keeping the open ecosystem relevant.

AI use at work is becoming normal, not exceptional

An Epoch AI/Ipsos survey of **2,021 U.S. respondents** found that among people who used AI in the past week, about half use it at least as much for work as for personal tasks [12, 13]. Among regular work AI users, **27%** said AI had replaced some tasks and **21%** said it had enabled new tasks [14]. Work use rose with paid access, from **38%** among free users to **76%** among employer-provided users; Microsoft Copilot was the most-used paid service for work, followed by ChatGPT and Gemini [15, 16].

Impact: The center of gravity is moving from casual experimentation to job workflows and employer-backed adoption.

Advanced cyber models are staying behind access controls

OpenAI clarified that the model being tested with a trusted tester group is a separate **cyber product**, not **Spud**, and that it is not being released publicly [17, 18]. Earlier reporting described a limited rollout to a small set of companies, similar to Anthropic's restricted cyber deployment pattern [19, 20].

Impact: For the most sensitive capability areas, frontier labs are moving toward staged, enterprise-style access rather than broad public release.

Research & Innovation

Why it matters: The most useful research this cycle focused on making agents more trainable, more reliable, and more efficient—not just bigger.

- **Atomic skills for coding agents:** A new approach breaks software work into five skills—code localization, code editing, unit-test generation, issue reproduction, and code review—and trains them jointly with RL. The reported gain is **18.7%** on unseen tasks like bug fixing and refactoring, without task-specific training [21].
- **Better verifiers for agents:** Microsoft's **Universal Verifier** separates process and outcome rewards, distinguishes controllable from uncontrollable failures, and manages long screenshot trajectories. It reportedly cuts false positives to near zero from **45%+** on WebVoyager and **22%+** on WebJudge [22].
- **Memory as experience: MIA** splits an agent into Manager, Planner, and Executor, with a loop of retrieve memory → plan → execute → store → improve. The authors report a new SOTA among memory agents, including a **+5.5** average gain and up to **+9.1** on harder tasks, with a **7B** model matching or beating larger closed models [23, 24].
- **Reasoning before post-training:** Meta FAIR's mid-training recipe adds interleaved thoughts, SFT mid-training, and RL mid-training before post-training. On base Llama-3-8B, the reported result is a **3.2x** improvement on reasoning benchmarks versus direct RL post-training [25].
- **Reality check for web agents: ClawBench** tests **153** live online tasks across shopping, booking, and job applications. Top models drop from around **70%** on sandbox benchmarks to as low as **6.5%** here [26].
- **Medical multimodal progress:** Google's **MedGemma 1.5** combines 3D radiology, whole-slide pathology, longitudinal X-ray analysis, and clinical document understanding in a single open-weight **4B** model. Reported gains include **+47% F1** in pathology and **+11%** in MRI classification over v1, and it outperforms Gemini 3.0 Flash on out-of-distribution CT analysis [27].
- **Cheaper inference research: Squeeze Evolve** reports up to ~**3x** API cost reduction and up to ~**10x** higher fixed-budget serving throughput across benchmarks including AIME 2025, GPQA-Diamond, ARC-AGI-V2, and MMMU-Pro [28].

Products & Launches

Why it matters: User-facing releases are increasingly about complete workflows—deployment, finance, healthcare, and visualization—not just chat.

- **LangChain Deep Agents deploy:** LangChain launched Deep Agents

deploy in beta as a model-agnostic, open-source agent harness for production deployment. It uses open conventions like **AGENTS.md** and **/skills**, deploys with short- and long-term memory on LangSmith, and exposes agents through **MCP**, **A2A**, and agent protocol [29, 30, 31].

- **Glass 5.5 API:** Glass Health released **Glass 5.5** via its developer API, saying it outperforms frontier models from OpenAI, Anthropic, and Google across **nine clinical accuracy benchmarks**. It also cut pricing by **70%** to **$3/1M input** and **$16/1M output** [32, 33].
- **Perplexity Computer + Plaid:** Perplexity's Computer now connects to Plaid so users can link bank accounts, credit cards, and loans, then track spending, build custom budget tools, and view net worth alongside investment portfolios. Computer tasks remain exclusive to Pro and Max subscribers [34, 35].
- **Gemini adds more interactive output:** Gemini can now create customizable interactive visualizations directly in chat, including adjustable variables, rotating 3D models, and data exploration [36]. Google also made longer **Lyria 3 Pro** music tracks available for free inside Gemini [37, 38].
- **Claude Cowork general availability:** Claude Cowork is now available on all paid plans, while enterprise customers get role-based access controls, group spend limits, usage analytics, and expanded OpenTelemetry support [39].

Industry Moves

Why it matters: New labs, enterprise partnerships, and infrastructure scale are defining where AI capability gets commercialized.

- **ElorianAI launches:** Former Brain/DeepMind researchers Andrew Dai, Yinfei Yang, and Seth launched **ElorianAI** as a multimodal reasoning lab focused on direct visual reasoning rather than translating images into text [40, 41, 42].
- **DatologyAI + Thomson Reuters:** DatologyAI said its legal-domain mid-training work with Thomson Reuters improved legal benchmarks by **5%** and general evaluations by **2.5%**, with **2.5x** amplification on Thomson Reuters' private legal evals using **<1%** of the original pre-training token budget [43].
- **Sandbox infrastructure is scaling fast:** A post on Modal's sandbox system said a major AI lab is already running about **100,000** concurrent sandboxes for RL workloads and aiming for **1 million**. Modal says it can spin up **hundreds per second** for a single customer [44].
- **AI in regulated services is attracting capital: Chapter**, which uses AI to help seniors navigate Medicare enrollment, reached a **$3 billion valuation** [45].

Policy & Regulation

Why it matters: Formal rules are still catching up, but release controls and governance warnings are already shaping deployment decisions.

- **Some frontier cyber systems are moving to controlled access:** OpenAI's cyber product is being tested with a trusted group rather than released publicly, and reporting compared its limited rollout to Anthropic's restricted cyber deployments [17, 19].
- **Demis Hassabis warns the next phase is harder to govern:** Hassabis said ChatGPT's launch locked labs into a "ferocious commercial pressure race" and warned that the coming "agentic era" in the next **2-4 years** will make alignment a much harder technical problem, calling for cooperation among labs, AI safety institutes, and academia [46].

  > "How do we make sure the guardrails are put in place so they do exactly what they've been told to do, and there's no way of them circumventing that or accidentally breaching those guardrails?" [46]

- **OpenAI's chief scientist is pointing to social fallout:** Jakub Pachocki said automating intellectual work raises major societal challenges around job displacement, wealth concentration, and governance of AI-controlled entities, and that these issues are coming faster than expected [47].

Quick Takes

Why it matters: Smaller updates still show where momentum is building across video, local AI, developer tools, and open-source agents.*

- **Muse Spark** reached **4th** in Text Arena, ahead of GPT-5.4 and Grok 4.2 [48].
- The **Meta AI app** climbed to **#6** in the App Store overnight [49].
- **HappyHorse-1.0** ranked **#1 or #2** across Artificial Analysis video leaderboards, with API access planned for April 30 [50].
- **YOLO26-MLX** brought native YOLO26 to Apple Silicon, with up to **2.6x** faster inference and **1.7x** faster training [51].
- **Hermes Agent** hit **#1 on GitHub Trending** and reached **40K stars in 45 days**, faster than OpenClaw's path to the same mark [52, 53].
- Anthropic's new **Monitor** tool lets Claude run background scripts that wake the agent only when needed; NousResearch said Hermes added a similar "notify when done" pattern three days earlier [54, 55].
- **Baseten BDN** promises **2–3x** faster cold starts for large models at scale [56].
- **Seedance 2.0** is now available to everyone on fal without restrictions [57].

Sources

1. X post by @OpenAI
2. X post by @OpenAI
3. X post by @OpenAI
4. X post by @wandb
5. X post by @claudeai
6. X post by @claudeai
7. X post by @alexalbert___
8. X post by @dl_weekly
9. X post by @GoogleDeepMind
10. X post by @UnslothAI
11. X post by @danielhanchen
12. X post by @EpochAIResearch
13. X post by @EpochAIResearch
14. X post by @EpochAIResearch
15. X post by @EpochAIResearch
16. X post by @EpochAIResearch
17. X post by @danshipper
18. X post by @kimmonismus
19. X post by @Techmeme
20. X post by @kimmonismus
21. X post by @dair_ai
22. X post by @omarsar0
23. X post by @TheTuringPost
24. X post by @TheTuringPost
25. X post by @jaseweston
26. X post by @arankomatsuzaki
27. X post by @kimmonismus
28. X post by @arankomatsuzaki
29. X post by @LangChain
30. X post by @LangChain
31. X post by @LangChain
32. X post by @GlassHealthHQ
33. X post by @GlassHealthHQ
34. X post by @perplexity_ai
35. X post by @perplexity_ai
36. X post by @GeminiApp
37. X post by @GeminiApp
38. X post by @GeminiApp
39. X post by @claudeai
40. X post by @AndrewDai
41. X post by @SethInternet
42. X post by @yinfeiy
43. X post by @datologyai
44. X post by @sarahcat21

45. X post by @steph_palazzolo
46. X post by @Ric_RTP
47. X post by @kimmonismus
48. X post by @scaling01
49. X post by @alexandr_wang
50. X post by @ArtificialAnlys
51. X post by @thewebAI
52. X post by @Teknium
53. X post by @chrysb
54. X post by @noahzweben
55. X post by @Teknium
56. X post by @baseten
57. X post by @fal