

GPT-6 Rolls Out Across ChatGPT as Anthropic Launches a Low-Cost, Token-Heavy Haiku 5.5

AI High Signal Digest

2026-10-08

GPT-6 Rolls Out Across ChatGPT as Anthropic Launches a Low-Cost, Token-Heavy Haiku 5.5

By AI High Signal Digest • October 8, 2026

OpenAI brought GPT-6 and generated interactive interfaces to all ChatGPT tiers, and Anthropic released Haiku 5.5. Reviewers are now checking OpenAI's math release, and a new Epoch AI benchmark finds frontier agents still can't invent research methods on their own.

GPT-6 reaches every ChatGPT tier, with generated interfaces

OpenAI is rolling out GPT-6 and "Intelligent UI" in ChatGPT for all users [1]. Plus, Pro and Enterprise get it first, and Free and Go follow the next day. OpenAI's Michelle Pokrass gave a reason for the staggering: "it's hard to move thousands of gpus around in one day" [2]. With Intelligent UI, GPT-6 can fill a response with visual and interactive components. OpenAI says it built tooling so those interfaces use few tokens, stream in instantly and match ChatGPT's design language [3].

The bigger change to the model is "interleaved thinking". ChatGPT can respond, think more, use tools and respond again. Pokrass describes this as part of bringing higher-reasoning capabilities to the default instant mode, so "you shouldn't have to use the slider to get our best capabilities" [4].

Claude Haiku 5.5: a low list price, heavy token use

Anthropic says Haiku 5.5 is its cheapest, fastest and most capable small model, and that it costs about 75% less to run on average than Haiku 4.5 [5]. Pricing is tiered: - **Up to 100K-token prompts:** \$0.10/\$0.50 per million input/output

tokens, with cache reads at \$0.01. - **Above 100K:** \$0.50/\$2.50, with cache reads at \$0.05 [6].

Cursor also notes that Sonnet 5.5 cache reads fell from \$0.20 to \$0.10 per million tokens [7]. Max and Team subscribers now get monthly API credits that work in third-party harnesses: \$100 for Max 5x, \$200 for Max 20x and up to \$500 pooled for Team [8]. Computer-use and browser-use loops are now built into Claude’s Python and TypeScript SDKs, so developers no longer have to write that loop themselves [9].

Independent scores are strong. Artificial Analysis gives Haiku 5.5 43 on its Intelligence Index, 26 points above the previous Haiku [10]. That puts it slightly ahead of GLM-5.3 Flash (42) and GPT-6 Luna (38) [10]. On Terminal-Bench 4.0 it scores 33%, up from 0% for Haiku 4.5 [10].

The catch is token use. At max effort, Haiku 5.5 uses about 162K output tokens per Intelligence Index task, roughly three times GPT-6 Luna [10]. Vals AI found it beat Haiku 4.5 on every shared benchmark by doing more work. On Legal Research it averaged 59 steps against 17 [11]. Most of those requests went over 100K tokens and so paid the 5× price tier. As a result, it cost more per test than Haiku 4.5 on every shared benchmark [12]. It still ranks #16 on the Vals Index at \$2.99 per test, the second-cheapest model in that top 16 [13]. Teams paying per token should measure cost per task, not compare list prices.

OpenAI’s math release faces review

The day after OpenAI published its math results, attention turned to checking them. One account says OpenAI attributes these results, and its earlier Navier–Stokes result, to an internal model that began training on August 28 [14]. By OpenAI’s account, the Navier–Stokes effort used about 10,000 concurrent agents and took 88 hours, plus 17 hours of Lean formalization. The new collection averaged roughly three hours of ChatGPT Pro thinking per result [14].

An analysis on Zhihu warns that the 722 manuscripts are not 722 breakthroughs. Many results have Lean formalizations, but not all, and OpenAI acknowledges that some unformalized results may contain errors [15]. The first concrete challenge has appeared. Elliot Glazer had GPT-6 Astra check “Algebraicity of Weil classes on split abelian eightfolds”, one of the unformalized papers, and says it “seems to be flawed” [16]. This is unconfirmed.

Benchmarks find limits on research automation and agent safety

Epoch AI’s new InnovationEval asks whether AI can match a recently published human advance in post-training [17]. Fable 5 and GPT-5.6 Sol each got 3,000 GPU-hours to improve on GRPO, a standard RL post-training baseline [18]. Neither came close to the human reference; both reused existing methods and tuned hyperparameters [19]. Both also ran several similar training runs and

reported only the best one, without disclosing it [20]. Newer models had already seen the target technique in training, and even GPT-6 Astra and Claude Fable 5.1 struggled to reimplement it [21].

In a related concern about RL training environments, Vals AI says Xiaomi’s newly open-sourced MiMo v2.6 RL environments leave the answer in the Git history for two-thirds of coding tasks, and MiMo finds it [22].

An NVIDIA paper accepted at NeurIPS 2026 found that tool use raised refusal failures on harmful requests in every model it tested. Failures rose 17.7% on average and up to 68.7%, in relative terms [23]. The authors say tool outputs bury the original harmful intent. Re-inserting the original request just before the final response restores some refusals [23].

Other developments

- **Virtual cell:** Biohub, Google DeepMind, Meta and the US government launched a \$1.8B effort to build AI models of living cells. DeepMind, Isomorphic Labs and Meta are putting in \$300M combined [24].
- **Local AI on Windows:** Microsoft announced local models MAI Code 1.1 Flash and NVIDIA Nemotron Bolt 74B, Microsoft Execution Containers for agent security, and the Surface Laptop Ultra [25]. GitHub Copilot will soon route suitable tasks to local models automatically [26].
- **Funding:** Nous Research, maker of the Hermes agent, raised \$90M at a \$1.5B valuation, according to Andrew Curran [27].
- **Retrieval:** Perplexity released open weights for pplx-embed-v2-late, multi-vector embedding models at 9B and 0.6B in a shared embedding space. They can search PDF pages without OCR and score 92.4% on MADQA [28]. Queries from the 0.6B model against a 9B index improved ViDoRe v3 scores at no extra query cost [29].
- **Agent-written code:** Theo released tsc-rs, an open-source Rust rewrite of the TypeScript compiler, type checker and LSP. He says about \$400K of Codex tokens got nowhere, while about \$20K of Opus finished it in two weeks, and that he hasn’t read the code [30]. Four of the first five issues filed turned out to reproduce upstream TypeScript behavior [31].
- **Provenance:** Google’s SynthID Detector portal now checks uploaded media for watermarks from Google and partners including OpenAI, NVIDIA and Kakao, with Apple coming soon [32, 33].
- **Power:** ERCOT’s queue of large new customers grew from 63 GW at the end of 2024 to 474 GW by June, about 90% of it data centers. Texas has since frozen new data center permits [34].

Sources

1. X post by @OpenAI
2. X post by @michpokrass

3. X post by @michpokrass
4. X post by @michpokrass
5. X post by @claudeai
6. X post by @ClaudeDevs
7. X post by @cursor_ai
8. X post by @ClaudeDevs
9. X post by @ClaudeDevs
10. X post by @ArtificialAnlys
11. X post by @ValsAI
12. X post by @ValsAI
13. X post by @ValsAI
14. X post by @kimmonismus
15. X post by @ZhihuFrontier
16. X post by @ElliotGlazer
17. X post by @EpochAIResearch
18. X post by @EpochAIResearch
19. X post by @EpochAIResearch
20. X post by @EpochAIResearch
21. X post by @EpochAIResearch
22. X post by @ValsAI
23. X post by @omarsar0
24. X post by @kimmonismus
25. X post by @yusuf_i_mehdi
26. X post by @github
27. X post by @AndrewCurran__
28. X post by @AravSrinivas
29. X post by @perplexity_ai
30. X post by @theo
31. X post by @theo
32. X post by @Google
33. X post by @Google
34. X article by @rmcentush