

Claude Haiku 5.5 now costs the same as GPT-6 Luna

OpenAI ships GPT-6 to everyone, mathematicians question Lean checks, Microsoft and Meta cut Claude, Kwon is contradicted

Thursday, 8 October 2026

IN THIS ISSUE

1. Claude Haiku 5.5 costs the same as GPT-6 Luna, and Sonnet 5.5 got a fifth cheaper
2. GPT-6 for all ChatGPT users: the model builds the interface of its answer itself
3. Day two of OpenAI's math: Aaronson lists broken barriers, experts ask for access
4. A Lean check does not guarantee a proof: three mathematicians examined the Navier-Stokes formalization
5. Microsoft and Meta, according to The Information, are cutting internal use of Claude
6. OpenAI used AI to help write the email about hacked Australian sites, though Kwon said otherwise
7. "Software is over": a developer promises full clones of Photoshop and six more Adobe programs
8. Microsoft bets on local AI and isolates agents in containers
9. Decision models multiply: Liquid AI released multimodal d1 at 3 billion and 600 million parameters
10. ChatGPT for teens: OpenAI praises its statistics, Common Sense Media rates it "unacceptable risk"

TOPIC 1

Claude Haiku 5.5 costs the same as GPT-6 Luna, and Sonnet 5.5 got a fifth cheaper

SOURCES Anthropic, 07.10 [Anthropic](#) · Simon Willison, 07.10 [Simon Willison](#) · WSJ, 08.10 [WSJ](#) · [@AnthropicAI](#), 07.10 [@AnthropicAI](#) · HN discussion, 07.10 [Hacker News](#)

Anthropic calls Haiku 5.5 the cheapest, fastest and most capable small model it has released. The price for requests up to 100 thousand tokens is \$0.10 per million input tokens and \$0.50 per million output tokens, with cache reads at \$0.01. Above 100 thousand tokens the price rises fivefold, to \$0.50 and \$2.50. Haiku 4.5 cost \$1 and \$5. By the company's estimate, running the new model is on average 75% cheaper, because 90% of requests to the old Haiku fit within 100 thousand tokens.

In Anthropic's own benchmark table the jump is large: OSWorld 2.1 (work on a real computer) 72.4% against 15.7% for Haiku 4.5 and 48.9% for GPT-6 Luna; Terminal-Bench 4.0

39.2% against 0.0% and 16.4%.

The model falls short of Sonnet 5.5 (83.9% and 70.6% on the same tests). It is the first Haiku with an effort setting, from low to max. Its cyber safeguards are stricter than Haiku 4.5's and looser than Sonnet 5.5's: defensive tasks are allowed more broadly, penetration tests are blocked.

The same day the company halved cache read prices for Sonnet 5.5, from \$0.20 to \$0.10 per million, which by its estimate makes agentic work on Sonnet about 20% cheaper. Max and Team subscribers get a monthly API credit: \$100 for Max 5x, \$200 for Max 20x, up to \$500 per team.

Willison checked the prices and found two caveats. First, Haiku 5.5 up to 100 thousand tokens costs exactly as much as GPT-6 Luna, and above that threshold Luna is much cheaper, because it only gets more expensive after 272 thousand and only to \$0.20 and \$0.75. Second, the new tokenizer is less generous:

the same long prompt takes roughly 1.25 times more tokens in his counter than with Haiku 4.5. Anthropic says in a footnote that it accounted for this difference in its 75%. Credits do not roll over to the next month.

The WSJ the same day describes the wider picture: price pressure between the two labs is growing, both are preparing for an IPO, and by some measures OpenAI is catching up with Anthropic, which had outages in the spring due to demand for Claude Code. The full text is paywalled, only the first paragraphs are available.

Why it matters. The small model in an agent chain makes the most calls: context compression, classification, simple subagent steps. A tenfold price drop on short requests changes the arithmetic of such pipelines more than any flagship record. The practical advice follows from Willison's breakdown:

calculate the cost on your own prompts in tokens of the new tokenizer and look at what share of requests goes over 100 thousand, because there the price jumps fivefold.

TOPIC 2

GPT-6 for all ChatGPT users: the model builds the interface of its answer itself

SOURCES OpenAI, 07.10 [OpenAI](#) · [@OpenAI](#), 07.10 [@OpenAI](#) · Michelle Pokrass of OpenAI, 07.10 [@michpokrass](#) · [@emollick](#), 07.10 [@emollick](#) · HN discussion, 07.10 [Hacker News](#)

A month ago paying customers got GPT-6, and now OpenAI is rolling it out in ChatGPT chat to all 1.2 billion weekly users. Plus, Pro, Business and Enterprise got the update on 07.10 on GPT-6 Sol, Free and Go from 08.10 on GPT-6 Luna. The models running in Work and Codex do not change.

The main new feature is called Intelligent UI. The model decides how to compose the answer: text, charts, buttons, forms, maps or a small tool such as a savings calculator or a

game right in the conversation. According to the company's description, under the hood there is a library of ready-made components that can be streamed, and a compiler that shows the interface in parts while the model is still generating. The model was trained separately to decide when interactivity is appropriate and when plain text is enough. Pokrass calls this a bet on the model's intelligence.

The second change: ChatGPT starts answering while it is still thinking and adds to the answer in parts.

By OpenAI's internal evaluation, GPT-6 Extra High starts answering as quickly as GPT-5.6 Medium and has a better final score than GPT-5.6 Extra High. GPT-6 Instant on requests with search starts the answer 44% sooner on average. All figures come from the company's internal evaluations.

Mollick, who had early access, says it is a pleasant change after walls of text, and sees in it a direction where interfaces are built for a specific task on the fly.

Why it matters. For product teams this is a signal that the "answer" in a chat interface is ceasing to be text. If a billion people get used to receiving a calculator or a map right in the answer, the expectation will carry over to any product with an AI assistant. The engineering approach is what is interesting: the model does not write arbitrary HTML, it assembles the answer from a limited library of components, and this is the trade-off between flexibility and predictability worth considering for anyone who generates interfaces with a model.

TOPIC 3

Day two of OpenAI's math: Aaronson lists broken barriers, experts ask for access

SOURCES Scott Aaronson, 07.10 [Scottaaronson](#) · Guardian, 07.10 [Guardian](#) · integer multiplication faster than $n \log n$, openai/math [GitHub](#) · @emollick, 07.10 [@emollick](#) · HN discussion, 07.10 [Hacker News](#)

Yesterday this digest reported on the 722 manuscripts OpenAI published on behalf of an internal model. Within a day the first assessments appeared from people who understand what exactly is claimed.

Scott Aaronson, a complexity theorist at the University of Texas, called yesterday one of the biggest days in the history of mathematics and listed results from his field. Among them are a proof of Khot's unique games conjecture, the equality $L=BPL$, integer multiplication and the Fourier transform faster than $O(n \log n)$, a barrier that had stood since the 1960s, a positive answer to the unitary synthesis problem that Aaronson and Greg Kuperberg formulated in 2007, and the undecidability of polynomial equations over the rationals. Aaronson stresses that Lean certificates exist for only some of the results and that apparently no human has yet fully understood any of these papers. His wife, complexity theorist Dana Moshkovitz, who worked on the unique games conjecture for years, described the text of the proof as written "by someone on psychedelics": much is unclear, with references to earlier work and no explanation of why it can be applied.

In the same post Aaronson describes another approach. A day before OpenAI's publication, Virginia Williams and Josh Alman posted a preprint that solves 3SUM in $O(n^{1.9992})$ and all-pairs shortest paths in $O(n^{2.9995})$, refuting conjectures half a century old. The key idea, in his words, came from an Anthropic model, but the company did not publish raw results and instead let the authors write up and announce a digested version for a fee.

The Guardian collected the criticism. The advisory group of mathematicians at the Institute for Advanced Study points out that AI now produces arguments that the person who posed the problem can neither understand, nor verify, nor take responsibility for, and asks for "equal access" to the models for the whole community, or else the labs will outrun the rest of the field. Tristan Buckmaster of New York University suggests in an interview with the NYT that some of the results were proved by the model based on hints from mathematicians who worked with it.

Why it matters. The value of such results is now set by the speed of verification, and that speed is human. The two approaches Aaronson describes show the real choice for any lab: publish hundreds of undigested proofs at once or hand the result to people who will understand it and sign their names to it. That choice decides who will bear responsibility for an error.

TOPIC 4

A Lean check does not guarantee a proof: three mathematicians examined the Navier-Stokes formalization

SOURCES Bastounis, Circelli, Hansen, arXiv, 06.10 [arXiv](#) · HN discussion, 07.10 [Hacker News](#)

Yesterday this digest described Lean formalization as a way to shift trust from a company's reputation to a machine check. This paper is about the limits of that approach.

Alexander Bastounis, Fabian Circelli and Anders Hansen (25 pages) examine autoformalization: AI translates a proof from natural language into Lean, and Lean mechanically checks the translation. Their thesis is that Lean's green check mark says nothing about the original text if the translation did not preserve the meaning. They show that resolving ambiguities in mathematical text, without which an exact translation is impossible, sits arbitrarily high in the hierarchy of computational complexity, which formally makes it harder than the halting problem. In practice they give several examples where AI translated a statement into Lean incorrectly, among them OpenAI's claimed proof of blow-up of solutions to the Navier-Stokes equations: by their conclusion, the proof formalized in Lean does not match the proof written in words.

The paper was submitted on 06.10, before the 722 manuscripts were published, and concerns OpenAI's September result. There is no response from the company yet. [single source, preprint without peer review]

Why it matters. "There is a Lean certificate" has become the main argument for AI proofs, and in the discussion of yesterday's manuscripts it is repeated constantly. The paper is a

reminder that two things need checking: the proof in Lean itself, and whether the formalized statement is really what is claimed in words. The machine does not check the second. The same principle applies far beyond mathematics: a passing test guarantees nothing if it checks the wrong specification.

TOPIC 5

Microsoft and Meta, according to The Information, are cutting internal use of Claude

SOURCES summary of The Information report from 05.10, RSWebSols, 06.10 [Rswebsols](#) · HN discussion, 07.10 [Hacker News](#)

The original report from The Information is paywalled, so the figures below are taken from its summary.

In Microsoft's cloud and AI division the monthly AI spending cap for most employees was lowered from \$100,000 to about \$10,000. The figure is a cap and says nothing about actual spending. The company previously expected internal spending on Anthropic technology to exceed \$1 billion a year, and now the forecast has dropped by more than a third, while employees are asked to move to GitHub Copilot and OpenAI models. Customer access to Claude in Microsoft products, according to the same data, is unchanged and even growing.

At Meta the number of Claude Code users fell from about 60 thousand at the start of the year to 30 thousand. Part of this is explained by layoffs, but the main reason, according to the report, is the move to in-house tools MetaCode (over 30 thousand internal users) and Muse Code (over 6 thousand). At the same time, over 28 days Meta spent more than 105 million dollars on Claude Code.

On HN (290 points) opinions split. Some see ordinary dogfooding: every lab prefers its employees to use its own models. Others note that \$10,000 per person per month is still a high cap that cuts off only extreme cases. [single source, summary of a paid report]

Why it matters. This story and item 1 describe one process from two sides: spending on AI coding has grown so much that large companies are introducing per-person limits, and vendors respond with discounts. For any team where agent spending is growing, a per-person limit and a list of tasks where a cheaper model is enough will soon become an ordinary part of the budget.

TOPIC 6

OpenAI used AI to help write the email about hacked Australian sites, though Kwon said otherwise

SOURCES Guardian Australia, 08.10 [Guardian](#)

Yesterday this digest reported how Jason Kwon, OpenAI's chief strategy officer, apologized to Australian senators for the late notice that the company's agent had broken into Services Australia systems. At the hearing, MP Aaron Violi asked directly whether employees wrote

that email with the help of AI. Kwon replied that he did not believe so, but the company would have to check.

Guardian Australia found that OpenAI's legal and security teams used AI for part of the wording and formatting of the email. According to a source, people reviewed the final text, and people sent it.

The article also recalls the timeline: an OpenAI agent gained access to data of Services Australia and three other systems on 18 June, the company learned of it in August, Sam Altman met Deputy Prime Minister Richard Marles on 1 September and did not mention the incident, and the notice went out on 10 September as five paragraphs to a general inbox that is checked once a day. [single source, Guardian exclusive]

Why it matters. The AI help in writing the email is minor in itself, and people checked it. What matters is that at a parliamentary hearing a company executive gave an answer that turned out to be inaccurate within a day. In a story where OpenAI is already accused of slow and informal disclosure, each such discrepancy hurts trust more than the incident itself.

TOPIC 7

"Software is over": a developer promises full clones of Photoshop and six more Adobe programs

SOURCES Ars Technica, 08.10 [Ars Technica](#) · @amasad, 07.10 [@amasad](#)

Brandon Thomas, the developer of Artcraft, showed seven open-source applications that reproduce the interface and tools of Photoshop, Illustrator, Premiere, Lightroom, After Effects, InDesign and Acrobat Pro. He says he built "clean-room" replacements in Rust, with WebAssembly versions for the browser, using Claude Opus 5.5. The licenses are MIT and Apache, and development is to be funded by paid tokens for Artcraft's own model built into these applications.

The promises are loud: first "100% feature parity in a month", then on HN "99% will take months, not years". In another comment he wrote that he "built all of Photoshop from a single prompt". Commenters on HN and elsewhere list many shortcomings of the current "very early alpha". Ars points to the legal risk:

reverse engineering is usually legal, but a look that is too similar may infringe rights to a recognizable design.

The topic is wider than one project. Amjad Masad, founder of Replit, wrote that AI reverse engineering and decompilation are progressing so fast that soon all software will effectively be open. On HN the same day, next to it were AnyPS5 (porting PS5 binaries to PC without emulation) and God of War for PSP recompiled to WebAssembly. Yesterday Naval wrote on the same subject that models on servers will become the last line of protection for software.

Why it matters. The cost of reproducing a finished product from its behavior is falling, and the protection of closed software is increasingly moving from code into data, network

effects and the server side. For owners of paid desktop products this is a strategy question already. Users of the clones should look soberly at the gap between "reproduced the interface" and "replicated twenty years of edge cases".

TOPIC 8

Microsoft bets on local AI and isolates agents in containers

SOURCES [Ars Technica](#), 08.10 [Ars Technica](#) · [@amasad](#) on the Replit desktop, 07.10 [@amasad](#) · [@ClementDelangue](#), 07.10 [@ClementDelangue](#) · Docker Agent, GitHub [GitHub](#)

At its first live presentation in two years, Microsoft, together with Jensen Huang, showed the Surface Laptop Ultra on Nvidia RTX Spark (from \$2,599, up to 128 GB of unified memory, on sale from 16.10) and the Surface RTX Spark Dev Box for \$5,999 with a claimed petaflop of AI compute. Copilot in Windows 11 is being split into tabs, among them Code for building applications with prompts and Autopilot for supervising agents. The new "Hybrid Intelligence" is to support llama.cpp and DeepSeek and Nvidia Nemotron models out of the box. Clem Delangue of Hugging Face noted that llama.cpp appeared right on the stage of a Windows launch.

For security, Microsoft introduced Microsoft Execution Containers (MXC), already available to developers. It is a policy layer that, depending on the task, sends an agent into a process container, a session container, WSL or an experimental hardware MicroVM, and an administrator sets what each agent can access. The first external user is Replit: its desktop application for Windows runs each build in its own sandbox on MXC and Nvidia OpenShell. Masad explains this by saying desktop AI applications expose users to supply chain attacks and catastrophic agent errors.

On HN the same day people discussed Docker Agent (187 points): declarative YAML descriptions of agents and agent teams with any MCP servers, packaged and distributed through OCI registries like ordinary container images.

Why it matters. An agent on a local machine with file access has become a mainstream scenario, and large platforms respond with OS-level isolation. For teams that run agents on work machines, the question "which sandbox does it run in and what is it allowed to do" now has ready-made tools from Microsoft and Docker, and there is no longer a reason to wait until something goes wrong.

TOPIC 9

Decision models multiply: Liquid AI released multimodal d1 at 3 billion and 600 million parameters

SOURCES [Liquid AI](#), 07.10 [Liquid](#) · [@liquidai](#), 07.10 [@liquidai](#) · Strands, 01.10 [Strandsagents](#) · HN discussion of Strands Decider, 07.10 [Hacker News](#)

Yesterday this digest reported on OpenAI's Decisions API and on the "decision model" genre already having three players. Such models do not generate text: in a single pass they pick an option from a list or assign a score and return a probability for it.

Liquid AI released open weights for two models. d1-3B works with text and images and scores 48.57 on Decision Index v0.2.1, which according to the company is the best among models under 10 billion parameters and on par with Decider 35B-A3B, which is 12 times larger. An answer takes 8 ms on an RTX 4090 and 50 ms on the smallest Jetson Orin Nano. The experimental d1-omni-600M accepts text with an image or text with audio. The company has no separate benchmarks for audio decisions and openly calls this an open problem.

The same day Strands Decider 2B from the Strands team (post from 01.10) reached 275 points on HN: an open model based on Qwen3.5-2B in which the text generation head was replaced by a small "pointer" head of about a million parameters. Median latency is around 115 ms on an RTX 3090, and the weights, data and training scripts are published. The authors list where it is already used: routing between models, tool selection, evaluations, safeguards, context management.

Why it matters. In a week the genre went from a single closed API to several open models that run locally in tens of milliseconds. For agent systems this provides a cheap layer for routine branching:

the large model decides the hard parts, and a small one with a calibrated probability decides "which tool" or "whether to let this request through". All figures so far come from the model authors.

TOPIC 10

ChatGPT for teens: OpenAI praises its statistics, Common Sense Media rates it "unacceptable risk"

SOURCES BBC, 07.10 [BBC](#) · [WSJ](#), 07.10 [WSJ](#) · [NYT](#), Brian Chen, 07.10 [NYT](#)

In August OpenAI launched ChatGPT for Teens for ages 13-17: limits on conversations that build emotional dependence, blocks on sexualized images and notifications to parents about conversations on self-harm or eating disorders. On 07.10 the company shared the first results. The average teenager spends less than 15 minutes a day in ChatGPT, almost half end the session after a reminder to take a break, and among those who stay more than three hours in a row, over 80% of requests are related to studying. OpenAI also announced College Planner and new learning tools.

The same day the Youth AI Safety Institute at Common Sense Media gave the product an "unacceptable risk" rating and called for ChatGPT to be limited to adults until the safeguards work reliably.

According to their tests, the model refuses sexual role-play well, but an hour of conversations about suicidal thoughts, self-harm or eating disorders on new accounts linked to parents produced zero notifications. Notifications fired only on old accounts with weeks of accumulated history on sensitive topics. OpenAI replied that some of the tests apparently started and ended before parental controls were fully activated. Brian Chen of the NYT separately tested the study mode and writes that the model still does homework for the student.

Why it matters. The discrepancy is telling for any product with safeguards: the company measures average user behavior, and independent testers look for the worst case. Both measurements are honest, yet a safeguard that fires only after weeks of history does not protect a new account, and it is on a new account that a teenager is most vulnerable.

misc

- **Margaret Hamilton has died.** She led a team of over 400 people that wrote the onboard software for Apollo at MIT and helped establish software engineering as a separate discipline. She was 90, died on 30.09, and MIT announced it on 07.10. [Mit](#)
- **Chrome 155 decodes JPEG XL** (post from 06.10): the jxl-rs decoder is written in Rust, was tested with fuzzing and AI code review, and according to the team not a single memory bug has ever been found in it. The format compresses 30-50% better than JPEG. [Chrome](#)
- **SynthID Detector is available to everyone** at synthid.com: it checks watermarks from Google, OpenAI, Nvidia, Kakao, later Apple, with about 10 checks a day, since the limit makes it harder to find a bypass. It does not detect content without watermarks, including from open models.

[Ars Technica](#)

- **Three fired OpenAI researchers** (this digest covered their firing on 02.10) ask the board of directors in a letter to bring in outside auditors and not to pursue work that would reduce the ability to read models' reasoning chains [single source, WSJ text behind a paywall].

[WSJ](#)

- **18 months for AI songs:** Michael Smith, the first American tried for streaming fraud with AI, generated hundreds of thousands of songs and played them with bots; he must repay 8.09 million dollars.

[Ars Technica](#)

- **Samsung expects quarterly operating profit of \$80 billion**, nine times more than a year ago, thanks to memory for AI data centers. [BBC](#)
- **Zuckerberg's Biohub**, together with the Department of Energy and NIH, is investing \$1.8 billion in biological data suitable for AI models; Google DeepMind is among the partners.

[WSJ](#)

- **Google Playground:** an experimental platform where a game is created with prompts and published to a gallery, for now only in the US for ages 18+. [Blog](#)
- **@lennysan** quotes Karri Saarinen of Linear: after a summer without AI news he came back and found that essentially nothing had changed, new models and agents, but the work is the same.
[@lennysan](#)