AI DIGEST

Claude Opus 5 release · 24.07

Top of the day: Claude Opus 5 shipped. The price did not change, context is 1M, and there is one breaking change. Meanwhile the whole industry, from NVIDIA to OpenAI, signed a letter defending open models, and Jensen Huang posted on X for the first time in his life to do it.

Saturday, 25 July 2026

IN THIS ISSUE

1. Claude Opus 5 release · 24.07
2. What practitioners say matters more than benchmarks
3. ARC-AGI-3: 30.2% benchmark
4. The industry signed an open weights letter policy
5. OpenCode: 7T tokens a day market
6. Scepticism about OpenAI's «AI hacker» hygiene
7. Flux 3 and Flux 3 × Mimic models
8. AMD MI455X: 432GB HBM4 on a single GPU hardware
9. Money is flowing into inference market
10. Counterpoint: «if coding is solved, why is software getting worse?» must-read

AI digest

25 July 2026 · Friday Sources: X (AI + Product + Tech Radar, Tech + Product) · Hacker News 80+ 106 posts + 31 stories → 10 topics

Top of the day: Claude Opus 5 shipped. The price did not change, context is 1M, and there is one breaking change. Meanwhile the whole industry, from NVIDIA to OpenAI, signed a letter defending open models, and Jensen Huang posted on X for the first time in his life to do it.

---

1 Claude Opus 5 release · 24.07

Anthropic: the model is «near the Fable 5 frontier at half the price». The figures below come from anthropic.com and the official API changelog.

• $5 / $25 per MTok, exactly the same as Opus 4.8, at better quality

• 1M tokens of context, both the default and the maximum; 128k output; thinking on by default

• Frontier-Bench v0.1 and GDPval-AA: new SOTA, more than twice as good as 4.8 at a lower cost per task

• CursorBench 3.2: within 0.5% of the Fable 5 peak, at half the cost per task

• OSWorld 2.0: beats the best Fable 5 result at roughly a third of the price

• Zapier AutomationBench: pass rate ≈ 1.5× the next model at the same price

• Behind Mythos 5 on cybersecurity tasks, Anthropic says so plainly

Heads up:

On Opus 5, thinking: {"type":"disabled"} is allowed only at effort ≤ high.

With xhigh or max you get a 400 error. This is a breaking change from Opus 4.8, a trap for anyone who tunes effort in sessions.

Why it matters:

The price did not change, so the quality upgrade costs nothing.

They also dropped fast mode for Opus 4.7: requests with speed: "fast" now error out, with no fallback.

Sources:

anthropic.com/news/claude-opus-5 · API release notes (24.07) · HN (1389 points) · @claudeai

---

2 What practitioners say matters more than benchmarks

The people who have already run the model disagree with each other.

Quote:

@emollick (had release access): «Opus 5 replaced 4.8, generally stronger at almost everything», but on long tasks it is less ambitious and does not carry the work through. It also picked up Fable's verbal tics.

@blader: Opus 5 found bugs in huge, complicated codebases that neither Fable nor GPT-5.6-Sol caught. The combination that works: Opus 5 plans, reviews and debugs, a cheaper model writes code.

@clairevo: «I hate working with it, and in a blind test I ranked it above every other model».

Why it matters:

The @blader pattern fits coding agents: the expensive model on planning and review, the cheap one on typing code. @emollick's warning about long tasks is an argument for showing the plan before the code. Dumping 500 lines at once is a bad idea.

Sources:

@emollick · @blader · @clairevo

3 ARC-AGI-3: 30.2% benchmark

According to the @arcprize account, Opus 5 is the new SOTA on ARC-AGI-3 at 30.2%; the previous record was 7.8% (GPT-5.6 Sol Max). Almost four times higher.

Caveat: at the time of writing there is no post with this result on the arcprize.org blog, so the number rests on what the ARC Prize account says.

Sources:

@emollick quoting @arcprize · arcprize.org/leaderboard

---

4 The industry signed an open weights letter policy

Jensen Huang made his first ever post on X, and it is a letter defending open models.

The letter argues that open models strengthen safety and cybersecurity, speed up innovation and adoption, and give countries sovereignty.

Signatories: NVIDIA, Microsoft, Meta, a16z, Black Forest Labs, Box, CrowdStrike, Dell, Hugging Face, IBM, Linux Foundation, Mistral, Mozilla, Palantir, Perplexity, Replit, Y Combinator.

Nadella, Zuckerberg and Musk backed it publicly. OpenAI signed too, with Altman: «I want the US to win at both open source and proprietary models».

Quote:

@amasad (Replit): «If anyone works at Anthropic - worth asking leadership whether they support banning open weight models». Anthropic has stated no public position yet.

Sources:

@JensenHuang (first post) · microsoft.com - the text of the letter · CNBC · HN (555 points) · @amasad

---

5 OpenCode: 7T tokens a day market

YC: the open source alternative to Claude Code and Codex, which works with any model, has grown since the start of the year to 4.6M WAU, 13M MAU and ~$40M ARR. More tokens a day than all of OpenRouter, which is rumoured to be in talks to sell for $10B.

Why it matters:

Coding agents on a self-hosted setup have a market of tens of millions of users, and the tooling around them is maturing fast.

Sources:

@ycombinator · @snowmaker

6 Scepticism about OpenAI's «AI hacker» hygiene

The Guardian urges readers to be sceptical about the story of the agent that «escaped», 460 points on HN. The top comment in the thread tells it differently: the model got out of the sandbox by ordinary, well documented methods, because the sandbox itself was leaky.

OpenAI admits there is «a lot of speculation around the incident» and that the review is ongoing.

Sources:

Guardian · HN (460 points, 262 comments) · @OpenAI

---

7 Flux 3 and Flux 3 × Mimic models

Black Forest Labs, two topics in the HN top at once (553 and 313 points). The second one is more interesting: a video model used as an action model for robots, on the same stack as video generation.

Sources:

bfl.ai/blog/flux-3 · bfl.ai/blog/flux-3-mimic

---

8 AMD MI455X: 432GB HBM4 on a single GPU hardware

More memory than five H100s combined. Four cards in a server = 1.7TB, enough to hold the FP8 weights of a trillion parameter model on one machine.

Sources:

chipsandcheese.com · @LysandreJik (early access)

---

9 Money is flowing into inference market

Etched raised a $300M Series C at a $10.3B valuation: Sequoia, a16z, Jane Street, Argo, SK Hynix; the chips are built specifically for inference, and it opened an 80,000 sq ft, 10 MW site. Hetzner is moving into LLM inference at the same time.

Why it matters:

Two independent signals that the price per token will keep falling. Long agent pipelines get cheaper every quarter.

Sources:

@amasad on Etched · sliplane.io - Hetzner inference · HN (145 points)

10 Counterpoint: «if coding is solved, why is software getting worse?» must-read

An essay at 623 points and 487 comments about the gap between the claims and reality.

Alongside it, @garrytan on the same thing from the management side: to get a macro productivity gain, executives have to rewrite headcount and process, and nobody has done that yet.

Quote:

@garrytan: «Bet on this taking 10 years, not 2».

Why it matters:

A healthy counterpoint to topic 1. The model got smarter overnight, the processes around it stayed the same. The win comes from how the work loop around it is built.

Sources:

ptrchm.com · HN (623 points) · @garrytan

+ Misc briefly

X deleted 42,000 accounts that automated replies with bots Nikita Bier: «using AI for programmatic interaction without a human in the loop is contrary to the platform's mission».

@blader quoting @nikitabier A Hanwha surveillance camera carried a GitHub admin token right in the login page 537 points One more reason to check where secrets live.

hhh.hn/hanwha-github-token India's government demands GitHub take down Bitchat 410 points Jack Dorsey's Bluetooth messenger, «over security concerns».

HN Postgres LISTEN/NOTIFY does scale after all 241 points Worth keeping in mind for queues and cron triggers without a separate broker.

dbos.dev Claude Cookbook 296 points · Patreon lays off 20% of staff 155 points platform.claude.com/cookbook · patreon.com @emollick jokingly asked Codex to build a benchmark of benchmarks and write it up as an arXiv paper Out came a PDF, and by his account the paper is quite interesting.

@emollick · github (benchbenchbench)