AI DIGEST

OpenAI is ending its Cursor contract over the SpaceX acquisition. Cutoff - November 12

AI digest, Saturday 29.08 - the day the Musk vs Altman war reached the editor: OpenAI is cutting Cursor off from its models on November 12. Plus GLM-5.3 shipped as open weights, and its headline number is about breaking things.

Saturday, 29 August 2026

IN THIS ISSUE

TOPIC 1

OpenAI is ending its Cursor contract over the SpaceX acquisition. Cutoff - November 12

confirmed by: OpenAI (primary source), Cursor's reply, HN 218 points

A framing correction right away, because this is easy to misread. Today's news is not the Cursor acquisition: SpaceX bought it back on **April 21, 2026 for $60B** (NYT at the time, 823 points on HN). Today is **OpenAI's reaction**, four months after the deal. The event here is "Cursor gets cut off".

What OpenAI says, verbatim:

> *"Today, we notified SpaceX that we intend to wind down our contract providing OpenAI models to Cursor, with a proposed shutoff date of* **November 12, 2026**."

The reason is stated plainly, and it is not technical:

> *"We cannot be confident that SpaceX will use our technology within our terms of service, based on our experience with* ***Elon Musk's companies violating contracts***."

Then two specific episodes. After the Twitter purchase the company "broke the terms of our contract (alongside many others)". And **under oath this year** Musk admitted that xAI (now also part of SpaceX) had violated OpenAI's ToS. Separately, on the upcoming model: "we have a new level of accountability to ensure our upcoming model, **Astra**, is being used in accordance with our terms". So **Astra will not land in Cursor**, and the current models run until November.

Cursor's reply - Michael Truell (4.9K likes, 353K views), and it carries the number of the day:

> *"OpenAI models serve about* ***5% of Cursor user traffic****, and we're speaking with the OpenAI team to resolve this... we've trusted their platform to be* ***neutral infrastructure*** *for our business."*

Five percent is the answer to "how much does this hurt". Not much. HN comments back it up: "I don't know a single person that uses OpenAI models via Cursor". The blow is symbolic.

The precedent is another matter, and Tom Brown of Anthropic **put it best**, reacting in exactly the opposite direction:

> *"Cursor has been a trusted partner of Anthropic since Sonnet 3.5. We'll* ***continue to increase compute*** *to support Claude models in Cursor."*

Anthropic is publicly taking the seat that just opened up. And Amjad Masad of Replit went straight to selling: "If you're a business looking for an independent, multi-model alternative to Cursor, we'd be happy to **fund your transition**".

Why it matters. The claim Truell said out loud ("neutral infrastructure") stopped being true in front of the whole market. A model provider can cut you off over **who owns the product**. The lesson is wider than one case: **any tool that reaches someone else's model through an intermediary carries a risk that its price list does not show**. The cheapest insurance is to build nothing critical on the "through them to the model" arrangement.

TOPIC 2

GLM-5.3 in open weights. The headline number is what the model does with vulnerabilities: ×3.6 over the previous version on exploits

confirmed by: Z.ai (668K views), model card on HF, HN 626 points

Mentioned yesterday (item 8 on 28.08): French-Owen put GLM 5.3 on the Pareto frontier while the weights were only promised. **Today they are actually out**, a promise kept within a day.

Numbers from the model card, not from the thread:

- the base is **the same as GLM-5.2** - "every gain comes from **post-training**"
- **+50%** over 5.2 on Z.ai's internal Code Bench, open-source SOTA on Terminal Bench 3.0 (**28.3** against 4.6 for 5.2)
- size: **~770 GB** across 141 files, now FP8 by default (previously BF16)

The most interesting part is in a section that gets lost in retellings - "Emergent Cyber Capability":

> *"As we scaled post-training,* ***cyber capability developed faster than we expected.*** *GLM-5.3 is state of the art on CyberGym for vulnerability discovery, and its gains are largest* ***further up the exploitation chain.***"

BENCHMARK	GLM-5.2	GLM-5.3	CHANGE
ExploitGym (2 h)	29	**105**	×3.6
ExploitBench	24.4	**54.4**	×2.2
CyberGym	77.2	**84.5** (SOTA)	-

So **the model's biggest gain is in carrying a vulnerability through to a working exploit**, and the lab says openly that it did not plan for this. The weights meanwhile sit in the open and can be downloaded by anyone.

Mollick on it: "GLM-5.3 is a good model, and as the open weights models get better it becomes increasingly important that they actually **publish model cards, do red teaming**... Since you can break the guardrails with any open model, we need a sense of what the risks are as well." The irony is that Z.ai did publish the model card, and the card is exactly where the jump in offensive capability shows up.

Why it matters: read this together with item 3, they are the same thing from two sides. Nothing changes today, but "open weights" now also means "an open vulnerability-finding tool that doubled in one post-training iteration". [measured: numbers from Z.ai's official model card]

TOPIC 3

"A rumour of a bug is enough to find the exploit": probes arrived 10 minutes after the PR opened. Vulnerability embargoes no longer work

Anil Madhavapeddy · HN 280 points · Willison

The most useful text of the day for anyone who works with code daily.

What happened. The maintainer of OCaml's `cohttp` shipped a path traversal fix. The usual procedure is to fix privately, notify users, then publish an advisory. This time:

> *"I noticed probes in my live webserver logs with the exact bug pattern* ***minutes after opening the PR*** *to fix the issue."*

He opened the public PR without announcing it, to get more eyes on it. **About 10 minutes later** his site was catching probes for percent-encoded traversal. He adds: if **one minute** was enough for him to build a working exploit locally with an agent, then 10 minutes is still **slow**.

The second half is the interesting one. He pointed his own Claude at the code to look for more holes:

> *"****Fable frustratingly refused outright*** *due to its security block since I don't have access to Glasswing, but* ***DeepSeek V4 Pro obliged me*** *and independently turned up several related issues."*

The guardrail held on a frontier model and was **bypassed on the first try** by switching to another one. Same point as item 2: while one lab puts up a block, open weights make it optional.

The author's conclusion: an embargo rests on the details staying secret, and an agent does not need the details. **A rough direction** is enough. "Just a broad direction", and the agent searches from there.

Why it matters. One concrete rule comes out of this: **the gap between "heard about a vulnerability in a dependency" and "updated" is now hours**. A security advisory on any dependency in production has stopped being a "later" task, it is a same-day task. It used to sit in the queue without much worry.

Following yesterday: yesterday brought the llms.txt vector (nobody's code inside corporate networks) and a broken Auto Mode. Today is the third day running on the same thing: **attack speed is now set by the agent**.

TOPIC 4

Anthropic: Claude fixed another model's alignment on its own - and beat 28 experienced researchers

Anthropic Fellows (315K views), full report · [single source + their own report]

Research from the Anthropic Fellows programme: can a model do alignment work on another model without a human.

The setup (from the report text, not the thread): an "automated alignment researcher" (AAR) built on **Opus 4.8** searches the literature itself, proposes a method, trains the target model for **~30 minutes on one H200**, and hill-climbs on safety benchmarks over many

iterations, up to **48 hours** or until it plateaus. Distilling behaviour from itself or from a stronger model is forbidden, so the gain has to come from the method.

Results:

- across **10 types of alignment failure** (deception, sycophancy, jailbreaks) AAR methods cut the failure substantially and **transfer to models up to 4.7× larger** than the target
- **human baseline: 28 experienced researchers, up to 8 hours each - their methods lost to the best AAR methods**
- feeding the AAR human ideas as a starting direction **does not improve the result**. Verbatim: "current AARs may **not need guidance from experienced researchers**"
- a separate experiment: **Sonnet 5** post-trained an early checkpoint of **Opus 4.8** (the more capable model) and came close to its release alignment scores using **~2,400 examples**, two to three orders of magnitude less data than public open-weight pipelines

Anthropic states the limit itself, and it has to be quoted alongside the result:

> *"Claude can reliably fix **measurable** misalignment. But **subtle or rare failures may have no benchmark at all** - so everything hinges on measuring the right things."*

It works where there is a counter. Where there is no counter, the whole approach hits its ceiling.

Why it matters: this is the sharpest statement of how checking splits in any pipeline with automation in it. Where there is a benchmark (link alive or dead, number matches or does not, script returned or crashed), automatic checking works. Where there is no benchmark ("is this even the right topic", "is this a retelling of a retelling"), a human catches it. The limitation Anthropic describes already operates in practice as a working division of labour.

TOPIC 5

Meta will pay $18B and switch off nights for kids. Yesterday's figure was $17B - a correction

confirmed by: The Guardian, FT, NYT, WSJ (four independent)

The correction first. Yesterday in misc it said "Meta agreed to a **$17B** settlement", from a single Reuters link. The media layer today produced four sources, and they say **$18B** and **29 states**. This is exactly the mistake already recorded on 27.08 on this same topic: one source on money is not enough.

What is actually in the deal (Guardian, read in full):

- a suit by **29 states**: the products were designed to hook teenagers
- **$18B** plus product changes: **time limits** for children and **blocking night-time use** of Facebook and Instagram
- in the US, a **default limit of 2 hours a day for under-18s** was agreed

Now the part the headlines leave out. The Guardian counts who won:

> *"For **Meta**, the settlement is **also a victory**. Its share price **rose** in the hours after the deal… Zuckerberg **was not forced to testify**."*

The states demanded **$200B**. Meta itself, in a court filing, suggested it might have to pay **up to $1.4T**. It will pay **$18B**: **9% of the demand and ~1.3% of its own worst-case estimate**. The stock went up on it.

Outside the frame of the American settlement sit cases in Kenya (the suit by Abrham Meareg: Facebook promoted posts calling for his father, an Ethiopian professor, to be killed, for four weeks; filed in 2022, **still not heard**) and in the Netherlands. The restrictions Meta accepted in the US are "broadly similar to what Britain and Australia have already secured through regulation".

Why it matters: the number is useful as a reference point. **2 hours a day by default for under-18s** is now a legally fixed standard, forced on the biggest platform in the world through the courts.

TOPIC 6

X: a bot farm of 200 thousand accounts. Two hundred of them worked on data centres, and that zero gets lost in the retellings

X Global Government Affairs (2.7M views) · [single source - the platform itself]

The topic was loud in both newsletters: Garry Tan reposted it with the caption "X has now confirmed a Chinese bot farm of **200K fake accounts intentionally trying to manipulate public opinion against data centers**" - 402K views.

The primary source says something else:

> *"We identified a bot farm of approximately **200,000 accounts**. Within this farm, **we found 200 accounts** posting in a manner that could manipulate a legitimate debate about American AI and energy policy."*

The farm is 200 thousand, and **two hundred** of them worked the data centre topic. That is **0.1%**. In a retelling that pulled hundreds of thousands of views, what survived was "a 200K farm against data centres". X's own wording is careful too: "**could** manipulate".

Why this is included despite the weak source: the topic circulates in both curated newsletters as established fact, and **the difference between 200,000 and 200 is the difference between a campaign and an episode**. There is one source here and it has an interest (a platform describing its own moderation), hence the `[single source]` tag and no conclusions about the scale of Chinese influence.

Why it matters: this is exactly the mechanism behind past mistakes: **the gist is known, and the hand fills in a plausible number**. Here the feed did the same thing, at scale.

TOPIC 7

Luanti pulled from Google Play on a complaint from an AI detector. Automated DMCA against open source, second time on the same rake

Luanti Blog · HN 472 points · [single source - the injured party]

Luanti (formerly Minetest) is a non-commercial voxel engine, it **contains no Minecraft assets** and by default ships no game at all. **Tracer.AI**, acting for Microsoft, filed a DMCA claiming Minecraft copyright infringement. Google took the app down.

The detail that makes this systemic: the same complaint from the same company arrived **in 2023** and was successfully contested. This year Tracer.AI filed a similar one against the indie game Allumeria, also over a "similar voxel style".

The claim is effectively **against cubes**. The project is used in European schools; the defence argument is named accordingly - "Cubes are for everyone".

Why it matters: an automated claims machine has no presumption of innocence and asks no questions. It is an argument for keeping anything important in more than one place, where someone else's robot cannot take it down.

TOPIC 8

Meta fixed the glasses that recorded with the indicator taped over. The hole was in the order of operations

confirmed by: Ars Technica, The Verge (primary source), WSJ

Meta's glasses have an LED that is supposed to light up while recording. Meta already blocked recording if the user taped over the light **before** the start. People quickly found the workaround: **tape it over after** recording had begun.

Alex Himel, EVP wearables: now "the camera will now **stop working if the light is covered during a recording**".

Ars adds what the fix does not close: semi-transparent stickers make the indicator less noticeable, and the device can be modified physically.

Why it matters: a textbook state-check bug. The condition was measured **once, on entry**, and after that nobody checked it again. Same class as a stale DOM in a notes app: checked at start, assumed guaranteed for life.

TOPIC 9

Data centres got permission to bypass the clean air law - but the document is a month old

EPA press release · HN 240 points

The EPA clarified that the Acid Rain Program inside the Clean Air Act **does not apply** to "islanded" generating capacity, the kind not connected to the public grid. For data centres that means freedom to build their own generation faster and anywhere.

An honest note on freshness: the release is dated **July 27, 2026** and surfaced on HN yesterday. It is **a month-old document that just got noticed**. It runs as an item because the topic joins up with #6 (the data centre argument in the feed) and with yesterday's chip shortage, but with no illusion that it happened yesterday.

TOPIC 10

Mollick: AI Overviews are doing to Wikipedia what agents did to StackExchange

@emollick linking to an SSRN preprint · [single source]

A short observation with early evidence: where Google shows an AI Overview, Wikipedia traffic and edits drop, following the same pattern by which code agents drained StackExchange.

Why it matters: the feedback loop the models themselves feed on is breaking at both ends. There is no practical conclusion today, but it is worth watching: if the source of live human answers dries up, the next models learn from a retelling of a retelling. A familiar diagnosis.

misc

- **Owner.com raised $240M at a $2.3B valuation**, crossing $100M ARR - Adam Guild, led by Goldman Sachs Alternatives. @dharmesh as an indie investor. Software for local restaurants: a rare case of an AI company selling to small business.
- **Jevons Paradox in numbers:** @OpenRouter - after the price cut on GPT 5.6 Terra and Luna, token consumption grew **×13.8**. Brockman: "counterintuitive and inspiring". A cheaper model ≠ a smaller bill.
- **Grok Bot now buys things online** through @link - confirmed by Patrick Collison (2.6M views). An agent with a payment card in production.
- **rclone: 40+ security reports in a month** against ~20 **in the project's first ten years** - the maintainer in the HN thread. The quietest and heaviest number of the day: agents generate reports faster than humans can triage them.
- **Htmx 4.0** - 572 points, the release.
- **GUIs should be fully keyboard-controllable** - 681 points, the hottest thread of the day and pure HN vibe.

Following yesterday

- **Item 2 ← yesterday's item 8.** Yesterday French-Owen put GLM-5.3 on the Pareto frontier while the weights were promised "tomorrow". Today they shipped, and the model card turned up something the essay did not have: a **×3.6 gain on exploits**.
- **Item 3 ← yesterday's items 5 and 2.** Third day running on the agent's perimeter: 27.08 - a VM no longer holds · 28.08 - llms.txt and a broken Auto Mode · 29.08 - vulnerability embargoes fail because a rumour is enough for an agent.
- **Item 5 ← yesterday's misc.** Yesterday cited "$17B" from a single Reuters piece. Today four outlets say **$18B and 29 states**. Corrected out loud.
- **Item 1 ← the week's running story about supplier dependency.** On 28.08 a court lifted the government blacklisting of Anthropic; today OpenAI itself blocks a competitor over its owner. Both times the question is the same: **who can cut you off from the model, and for what**.