AI DIGEST

Ten open problems, $2000 and a Lean certificate - and why Daniel Litt is the reason to trust it

All week the feed was about agents breaking out of sandboxes, and yesterday the theme closed with Tailscale's numbers. Today the feed swung the other way - and this is the biggest single-day spike in attention since this digest started: Noam Brown's post at 7.04M views. OpenAI said an internal version of the Astra model solved ten open problems in mathematics, each of which had stood for at least a decade. The main thing happened over the next 12 hours: a mathematician who argued publicly and had a bet on it lost that bet and said so.

Sunday, 2 August 2026

IN THIS ISSUE

1. Ten open problems, $2000 and a Lean certificate - and why Daniel Litt is the reason to trust it 422 points on HN / 287 comments; Noam Brown (@polynoamial) at **7,040,000 views and 13,827 likes**, Brockman at 682K. By reach this outweighs every other topic of the week combined.

What the **primary source** actually says (the OpenAI page, not a retelling in a tweet):

• the results came from "**an internal version of Astra, our next large model**"

• the problems are ones where "**there had been no progress on the main result for at least a decade**, and in most cases far longer"

• the token cost of finding all the solutions was "**about $2,000 at Sol API rates**"

• the manuscripts were prepared by **people together with the model**, after which the model **formalised every argument into a Lean certificate** - the proofs are machine-checked

The list is specific: sphere packing in high dimensions, bounds for binary codes, **the existence of non-sofic groups**, **a refutation of Connes' rigidity conjecture**, lower bounds for the permanent, quantum parallel repetition, the closest vector problem in lattices (straight into post-quantum cryptography), Ehrhart's volume conjecture, multicolour Ramsey numbers (**Erdős problem 183**) and two more Erdős problems, **146 and 180**.

Now the reason this is item 1. A lab's claims stay marketing until someone with no stake in them confirms it. Ethan Mollick was waiting for "the verdict of one of the most sober and AI-informed mathematicians" - that is **Daniel Litt (@littmath)**, an actual professor of mathematics. His verdict: "**It's a big deal**" (102.8K views). Three hours later he wrote something that weighs more than any benchmark:

> "***I concede the bet.*** *Strictly speaking it was not resolved - it seems an Annals-level number theory paper has not appeared yet - but it is clear that* ***my expectations about the capabilities needed for such a paper were wrong****, and it is only a matter of time." And then: "I should have chosen a better proxy for the kind of research I care about most. Maybe that is a lesson for the profession too."*

He also drew the line on his own competence: "most of this is far outside my competence (I know the most about sphere packing, and I like it a lot, but **only as an interested amateur**)".

Why it matters. Three independent things landed at once: a claim with a specific price attached, **machine verification** (Lean - a proof either compiles or it does not), and **a named sceptic publicly surrendering**. When only the first one is present, it is a press release. Today all three are there, which is why this is filed as an event. [proven - as to the publication and the Lean formalisation; **fuzzy** - as to how much of this survives review by the mathematical community; Litt himself says the key bar has not been cleared]

Noam Brown · Litt's verdict · conceding the bet · HN thread

2. Cursor removed prices from its interface - the exact metric harnesses were measured on two days running 317 points on HN, 143 comments. A thread on Cursor's own forum.

The facts: on **31.07** Cursor removed the **dollar cost** from the usage page and from the CSV export for self-serve plans (individual and Teams). The **Spend** metric is gone, so is the **Cost** column with the per-request price, and so are the dollars in the export - it now shows `$0.00` or nothing. The official explanation from a staff member (Kevin Neilson): this was "**a deliberate decision**" to reduce confusion - "anything covered by the plan shows a token count and is marked *Included*, because nothing is charged for it".

Users did not buy it. Verbatim from the thread: "That window was used to keep a close eye on daily spend"; "That dollar figure read as the value (the savings) I was getting"; a Teams admin with **$30K** of spend: "How do I track spend per user now, the way I used to?".

Why it matters. On Thursday there was the Composio benchmark - a **3.7× price difference between harnesses** for the same result; yesterday Maleev gave the corporate frame for token-maxing. Today one of the largest harnesses **removes the very counter** that such a check runs on. The conclusion is narrow: cost has to be measured on your own side, because someone else's UI can drop it on a product manager's say-so. Your own measurement (how many kilobytes in how many seconds) is the right form of ownership over a metric. [proven - this is Cursor's official answer in their own thread]

Cursor forum · HN

3. Karpathy: Lord of the Rings for 1M tokens (~$10), 5500 lines of code and 2 hours - and an honest conclusion about what LLMs cannot do 283K views in an hour, 4,947 likes - fresh, it landed at 06:00 Kyiv time.

The experiment, verbatim: Opus 5 was given **the first paragraph of Lord of the Rings**, a budget of **1M tokens (~$10)** and the job of making a three.js render. "Opus went off for **~2 hours** and wrote **5500 lines of code** that (procedurally) rendered the story. Looks a bit janky but it's fun."

Two thoughts worth more than the demo clip itself:

• "**No sane person would ever spend the time** on something this bespoke, but an LLM has all the stamina and patience in the world - so you go from *'nobody would do this'* to *'sure, why not, it's nearly free'*."

• And immediately, where it breaks: "the worlds/games domain **exposes a weakness of LLMs: they cannot easily verify their own work**, because they cannot efficiently and natively perceive video or play what they made".

Why it matters. The second thought describes a general class of problem. A typical case: a script hangs, the cause is hunted in the network for several minutes, because the process sitting at 99% CPU **is not visible** - the thing a human would spot at once. A model without the right perception channel will confidently do the wrong thing. The practical takeaway: wherever a result is marked "done" but **cannot be checked through the same channel a human would use**, that is where an explicit check belongs. The first thought works too: a whole class of one-off homemade things that never made sense before just got cheap. [proven - a self-report with concrete numbers]

Karpathy

4. QM in one day: 2,475 → 5,234 stars, and inside YC people use it daily A continuation of yesterday's item 2 - the number doubled.

Yesterday it was **2,475 stars**; the GitHub API now returns **5,234 stars and 538 forks** (repo created 29.07, last push 01.08). That is **+2,759 in a day**.

New in the last day: accounts from inside. Jon Xu (@xuster): "uses QM **every day**, and it has become an inseparable part of the work". And a detail that was not there yesterday (Anna Zhang, @anna_y_zhang): "The coolest thing about qm is **the culture it grew out of**: YC once described how their **agent sessions are streamed inside the company**, so people learn by watching how others work with agents. **An organisation's agent history as shared learning material.**"

Why it matters. That last part is a cheap idea for any agent stack that already stores session logs. Such logs usually get mined for knowledge. *How the work is done* gets lost. The cheapest step: once a week pull two or three places from the log where the wrong path was taken and file them in the knowledge base as a "how not to" line. No code changes, just a change in what the consolidation looks for. [proven - GitHub API numbers, verified]

QM repo · Jon Xu · Anna Zhang

5. "AI does not make working products - that is still human work" 250 points on HN, **267 comments** - one of the hottest discussions of the day.

The author's thesis: AI sharply shortened the path to **the first working version**, but the distance to production stayed the same - system design, scaling, error handling and security sit exactly where they were. The strongest lines: "**Hard problems in software were never about syntax. They were about judgement.**" And: "Models have no judgement. They have pattern matching **with an eagerness to produce code they believe matches the intent**." The conclusion: "Learn the fundamentals first. Then the new tools. **In that order.**"

Why it matters. This is the antidote to item 3, which is why they sit together. Karpathy shows a model can grind out 5500 lines over two hours; this author points out that **the distance from there to a working thing has not changed**. A typical example is a performance bug where the fix is one line, and all the value is in measuring the thresholds (fine at 4 KB, hangs at 8 KB) and working out that the cost is quadratic. A model that "eagerly produces code matching the intent" would have written a workaround and moved on. [promising - a practitioner's argued essay, not research]

article · HN

6. Mollick: AI is blurring the lines between professions - and this is the second independent confirmation 136K views. Formally a continuation of yesterday's misc, but today it gained weight and has **two independent sources**, so it gets its own item.

Mollick: "One of the main findings of the study at **Procter & Gamble** was that **AI blurred the boundaries between jobs**. Now **OpenAI has a similar result**. Organisational boundaries become permeable, the walls get thinner." And his harder phrasing: "This is **not optional**, everything gets messy, and ignoring the change will not make it go away."

What it looks like in practice (his own example): "If a marketing project needed a bit of code, or a financial model in Excel, or a solution to a sphere-packing problem in high dimensions, you used to have to call the engineering department, the financial analysts or the theorists in the maths department. **Now it is just a question for the AI.**" Alongside him,

independently, Lenny Rachitsky: "**Everyone is becoming a part-time engineer and a part-time marketer.**"

Why it matters. Work is already built this way: one person reaches into the creative, the data and the code. The flip side is in Mollick's own words: boundaries become permeable **in both directions**, so the defence against blurring is having clearly fixed quality criteria where the expertise is missing. [promising - a peer-reviewed P&G study plus an independent OpenAI report, but this is organisational sociology, not measurement]

Mollick · P&G paper · Lenny

7. Mollick separately: capability gains are becoming impossible to "feel" 162K views - the soberest thought of the day, which is why it gets its own item.

Verbatim: "For nearly every human on the planet this is beyond comprehension. You can only trust expert mathematicians when they tell you whether it is impressive. **This is starting to happen across many fields, and it makes capability gains harder and harder to "feel".**"

Next to it: "The point is still the same: **the lack of verifiable answers** in many fields is a real problem for LLMs, but not as big a one as it is sometimes made out to be." And an honest counterexample from him: "**They are still bad at good long-form fiction.**"

Why it matters. If gains cannot be felt directly, the only substitute for the feeling is **outside verification**: the Lean certificate in item 1, a named sceptic conceding in public, an independent practitioner instead of a lab announcement. Any news about new model capabilities is worth reading by the same principle. [proven - his observation, not a claim of measurement]

Mollick

8. Levie: the gap between "AI in everyday life" and "AI in deep domains" will start to widen 72K views. Aaron Levie (Box) is the same voice who yesterday framed the harness as the main variable.

His thesis: "You will see **a growing divergence** between what AI does in personal life and everyday productivity, and what it can do in very deep domains - maths, science, law, code. Up to a point the growth in capability felt **even across all domains**..." - and then comes exactly the break the Astra story showed today.

Why it matters. A practical takeaway for expectations: "solved ten open problems" and "remembered a small household errand" are **different curves**, and the first does not pull the second. So it is no surprise when a model that crushes theoretical computer science fumbles a routine chore, and a fumbled chore says nothing about the ceiling. [fuzzy - a practitioner's forecast, not data]

Levie

9. Cursor is not alone: 456 points for a breakdown of how Google buried RSS - and this is an anniversary Top score of the day on HN (456 / 159 comments). **The article is from 2023**,

it just resurfaced, so it is not today's event.

The timeline from the text: Chrome removed the built-in RSS subscribe button back in the early 2000s with no explanation; **2007** - Google buys FeedBurner, **October 2012** - shuts off its API, **July 2022** - drops most of its services along with email subscriptions; **2013** - Google Reader closes (an engineer from inside: "the whole time I was on the project, various people were trying to kill it"); **December 2017** - Google News removes RSS support entirely. The author calls this "**Embrace, Extend, Extinguish**".

Alongside it the same day: "**Has Google abandoned Google News?**" (261 points) and "**A directory of people who love RSS**" (169 points). That is the mood of the day.

Why it matters. For any pipeline that collects content automatically, RSS still carries the weight: a public feed does not break on an expired session and needs no login, unlike a social network API. That is exactly the durability HN is voting for today with 456 points. [proven - the article's facts; dated 2023, not today's event]

article (2023) · HN · Google News

10. The first YC company to hit 100K GitHub stars during a batch - the number checks out
Jared Friedman (YC): "**Graphify Labs is the first YC company to hit 100K stars on GitHub during the batch.**" 67K views.

Checking the API showed that claims like this are the easiest to inflate: **100,387 stars**, repo created **03.04.2026**. So 100K in four months, and the number is real. What it does, from the repo description: it turns any codebase **along with its docs, SQL schemas, configs and PDFs** into something you can query.

A curious detail nearby in the results: the third most-starred repo for the query "graphify" is `claude-code-memory-setup`, promising "**up to 71.5× fewer tokens per session** in Claude Code via Obsidian + Graphify" (906 stars).

Why it matters. The interesting part is not the tool but the measurement under it. If an agent's memory is already structured and searchable, how many tokens its injection eats in every session is usually something nobody has measured. That is a cheap check, worth doing before looking at anyone else's tooling. [proven - as to the 100,387 stars, verified through the GitHub API; **fuzzy** - as to the "71.5×", which is repo marketing, not an independent measurement]

Friedman · Graphify repo

Misc • Seedance 2.5 from ByteDance (225 points on HN) - a video model with "one-take creation" and flexible referencing. In the comments the main argument is about what it is for: "the only real use for these models is disinformation and spam" against "any director sees the potential here to make scenes that are otherwise impossible" seed.bytedance.com · HN

• **@levelsio published correlations from his own tracker**: he wired a calorie app to an Airthings air sensor and a WHOOP. Best for sleep is **a cold, dry bedroom**; the split: cool

≤18.1°C, middle 18.1 - 19.4°C. Top positive factors: low calories **+6.5%**, low fat share in food **+6.3%**, dry bedroom **+5.2%**, a bath **+3.0%**. He writes himself that "not all are significant, more data needed" - this is n=1 with no control, so treat it as a hypothesis post · on the bedroom

• **The same author on creatine and hair loss**: "Creatine causing baldness - debunked. You go bald from genetics, and it affects 80% of men". His argument: the age when men start going to the gym and taking creatine **coincides** with the age hair loss begins, so there is no causal link. He also takes a swing at finasteride/dutasteride as "synthetic 4-azasteroids". This is his opinion, not consensus 1 · 2

• **Creatine and depression** (184K views, same author): a review of five clinical studies in the Canadian Journal of Psychiatry - creatine was **added** to antidepressants or psychotherapy. Presented as a quote from someone else's post with no link to the primary source, so taken as [fuzzy] post

• **Brockman on ChatGPT in Slack** (585K views, 7,091 likes - top of the day after the maths): "At OpenAI a lot of people connect their ChatGPT to Slack. People **really dislike** it when a colleague's ChatGPT asks them for help - even when they would have been perfectly happy to do the same work if the colleague had asked in person." The sharpest illustration of why an agent should write under its own name and not pretend to be a person post

• **Cursor is not the only one being dropped**: in the same HN thread practitioners describe leaving Cursor en masse for Claude Code / Codex / Zed, and the most common argument is "Cursor as an IDE without the LLM features is just a worse VSCode"

• **Masad on an 8B model at chess**: Qwen-chess beats frontier models and Stockfish level 0 by **~1500 Elo**, spending **1-2 seconds per move against 30**. "Love seeing an 8b model take GPT 5.6 on high reasoning" post

• **McKay Wrigley under Karpathy's post**: "Someone please **fund the entire trilogy in this form as a benchmark**" post

• **Koomen (YC)**: the share of YC applications containing the word "wedge" went from **<1% in spring 2025 to >20% in summer 2026**. A good indicator of how fast jargon eats thinking post

• **NetBSD 11.0** released (237 points) HN · **Go 1.27 Interactive Tour** (86 points) HN

• **"Your development pipeline is a production system"** (161 points): the argument that a broken CI should be treated like a production outage. A precise line from the comments: "we grumbled about how dumb it was to cut costs on pre-prod... **because pre-prod was prod**" HN

• **Garry Tan on the 2026 shift**: "The most interesting vibe shift of 2026 - **OpenAI really does look like an open platform**: intelligence as a utility, against the signal that the optimal move is to integrate vertically across the whole stack" post

• **Lenny on systems thinking** - the same thesis as yesterday, with new confirmation: it came up in **five podcast interviews in a row** post

• **Flint from Microsoft** - a visualisation language "for the AI era" (257 points). In the comments the community answered in unison that **ggplot / the Grammar of Graphics already solved this**, and "for the AI era" is just the obligatory buzzword HN
