AI DIGEST

OpenAI published the first measured results for Jalapeño, its own inference chip. And SemiAnalysis, who handled it in the lab, immediately explained why the Blackwell comparison is unfair

AI digest, Wednesday 26.08 - the day hardware and the law moved faster than models: OpenAI showed its own chip with measured numbers, and X legally shut down every Twitter mirror.

Wednesday, 26 August 2026

IN THIS ISSUE

TOPIC 1

OpenAI published the first measured results for Jalapeño, its own inference chip. And SemiAnalysis, who handled it in the lab,

immediately explained why the Blackwell comparison is unfair

The biggest story of the day and the richest in detail. Three layers of sources, and they do not agree with each other.

Layer 1, OpenAI itself. A three-tweet announcement (25.08, 17:19, **1.3M views**) and the results page. `openai.com` is a **blind domain** (403 on both the real page and a deliberately invented one, checked back on 02.08), so it was read **in a logged-in browser**, not with curl. Verbatim from the page:

> *"Across all three, Jalapeño delivered* **1.5 to 1.9 *times more AI work per watt*** *at peak throughput and* **1.7 to 3.6 *times lower end-to-end latency*** *than the comparison systems. For highly interactive workloads, it delivered* **2.1 to 4.1 *times higher performance*.**"

A correction to the retelling, right away. @TheRundownAI wrote "up to 1.9x more work per unit of electricity… responses up to 3.6x faster". Formally true, but it **takes the top of the range and presents it as the number**. The primary source says **1.5 - 1.9×** and **1.7 - 3.6×**: the lower bound is nearly twice as modest. A classic reason to take numbers from the blog post.

Also from the primary source, on power draw: the chip is **rated at 700 W, but measured sustained power is 550 W and below**; results were normalised by each accelerator's *rated* power. The test models are public: GPT-OSS 120B, DeepSeek R1 670B, Kimi K2.5 1T. The benchmark is **InferenceX by SemiAnalysis**, someone else's.

Layer 2, SemiAnalysis, who were in the lab. Their analysis - **386 points on HN**, thread. Reposted by Elad Gil ("Impressive speed to tape out") and Greg Brockman. What they add beyond the press release:

- Development cycle: started mid-2024, tape-out **in ~16 months**. For an ASIC that is very fast, and they say it outright: "It shows that claims that use of AI is being used to accelerate chip design **are real**".
- The chip is **not specialised for OpenAI's models**, despite half the press writing that: "Jalapeño is a **generalized** inference chip". As a demo OpenAI ran Doom on it, ported **by Codex prompts alone**.
- On DeepSeek R1 - **over 700 tokens/s per user at concurrency 1**, and that is **without** speculative decoding and without prefill-decode disaggregation. On GPT-OSS ~1400 tok/s/user.

And three caveats from the same people, absent from the announcement:

1. **"All numbers are provided to us by OpenAI."** The InferenceX runs were verified in person in the lab, but **the full set was not run** and **AgentX results were not seen**, and AgentX is exactly long context and multi-step dialogue, meaning **real agentic load with cache behaviour**. Verbatim: "Frameworks that perform well on 8k1k may perform worse on AgentX".
2. **The Blackwell comparison is "somewhat incomplete and unfair"**: Jalapeño with HBM4 competes with **Rubin**. And Vera Rubin **is already shipping to customers**, while OpenAI

still has engineering samples.
3. The tested models are **not on the open frontier**: NVIDIA and AMD publish results on larger ones (DeepSeek V4 Pro, Kimi K3).

Rollout is "by year-end", Gen 2 in development.

Why it matters. Practical value today is zero: the chip is not for sale, a subscription stays a subscription. Two things are still worth noting. First, the **methodology**: this is a perfect view of the trust gradient inside a single topic. Press release, then a range trimmed to its upper bound in the retelling, then independent people who were in the lab and say "they gave us the numbers, the full set was not run". These are **three different levels of evidence**, and the difference is worth keeping out loud. Second, **why performance per watt**: SemiAnalysis say plainly that OpenAI is constrained by **data-centre power**, with budget to spare. That is the same line as yesterday's item 6 about the data-centre moratorium. Tag [promising]: the measurements are real, but the source is one-sided and there is no agentic benchmark.

TOPIC 2

X sent cease and desist letters to every Twitter mirror. Nitter and XCancel are down

725 points, the second story of the day.

The GitHub thread started plainly, with a complaint that "all public instances are not working". Thirty minutes later maintainer zedeus replied verbatim:

> *"We have received **cease and desist letters**. Awaiting legal advice at the moment, but for now expect **all nitter instances to remain down for the foreseeable future**."*

XCancel is a separate story at 223 points, and the text there comes **from their own site**: "On Monday **24th August at 8PM EST**, we received a letter from X Corp. asking to cease and desist… The service XCancel is stopped until further notice. Thank you for the trust you have put in these **two years** of XCancel".

One thing from the thread was checked, and it did not hold up. A commenter wrote "zedeus was forced to delete the repo" / "x corp forced zedeus to delete it". Checked through the API: `api.github.com/repos/zedeus/nitter` returns **200**, the repository is in place. Instances are down, but the code itself did not go anywhere, and those are different things. A classic case of an emotional comment in a hot thread reading as fact.

Also from the thread, practical: one instance operator writes that **self-hosting still works** on his own residential IP, and adds that XCancel got a letter too, but "XCancel is not in the U.S., so it might be more resilient".

Why it matters. The signal reads wider than the fate of two services: X is systematically closing every path to read it without an account, and the same thread mentions the clampdown on anonymous viewing and the verification requirement. For any pipeline that

collects posts, one option is left, a logged-in collector, and it stays alive exactly as long as its session does. That session will eventually die, and after today there will be nothing to replace it with: a pipeline with a single path has nothing to recover on. Worth knowing in advance. The day the collector returns zero posts is too late for that conclusion. Tag [proven]: a statement from the maintainer and from the service itself.

TOPIC 3

Ramp: 75% of merged PRs are written by their own agent, not Claude Code or Codex. And this is the third day running with Ramp in the issue, but the first time as engineering

Gergely Orosz's piece (25.08) - an interview with Ramp CTO Rahul Sengottuvelu, head of engineering Hamid Dadha and project founder Zach Bruggeman.

This is Ramp's third appearance in three days, and the angle is different each time: on 24.08 their payments data gave the item on Opus 5's share; on 25.08 that same anomaly resolved through ZDR; today it is how they build.

The numbers, and they are serious:

METRIC	VALUE
PRs authored by Inspect	**75%** of all merged (January - 60%, May - 75%)
Sessions all time	**over 1M** (milestone in July)
Internal contributors	**150+ people**
Sandbox spin-up time	**≤ 5 seconds**

And Ramp are not alone here: **Block has Goose** (open source), **Stripe has Minions**, **Shopify has River**. So this is a pattern.

Why they did not take something off the shelf (verbatim reasons from the text):

- **A local machine cannot run many agents in parallel.** "Liked Claude Code on day 1, but were constrained by only being able to run **one or two sessions** on local machines".
- **Verification.** Inspect runs tests, **watches telemetry and reads feature flags**, and the frontend is checked **with screenshots and live previews**. Orosz writes plainly that third-party harnesses cannot do this "out of the box", because they **have no access to internal telemetry and flags**. Ramp built screenshot verification "almost a year ago, before vendors supported it".
- The stack as it is: React/Vite, **Cloudflare Durable Objects**, SQLite, Cloudflare Agents SDK, **Modal** sandboxes; inside the sandbox - **OpenCode**, Postgres/Redis/RabbitMQ/Temporal, Chromium and VS Code Server.
- The social detail, which is the most interesting part here: **all Inspect sessions are public and open to collaboration, with no opt-out**.

Why it matters. The most practical item in the issue, because it is **about architecture**. Orosz's main thesis: "the only constraint on agents' ability is **model intelligence, not missing tools or access**". This maps onto any agentic pipeline almost word for word:

1. **Verification closes the loop.** Ramp won by making the agent **see the result of its own action**. The weak spot is typical: intermediate artifacts get checked better than the final one the reader actually reads. The 17.08 mess with raw markdown and the 02.08 one with the PDF are exactly that.
2. **"Buy, don't build" on harnesses** is the conclusion of a company that could buy anything. The reason is specific: **integrations with internal context**, which a vendor will never have.
3. **Parallelism runs into hardware, with the model unchanged.** They solved it with remote sandboxes. A single machine sets the ceiling as soon as you want several heavy sessions at once.

Tag [proven]: numbers from the company itself in a named interview, but this is **their own data about themselves**, with no independent measurement.

TOPIC 4

Laude/MIT released Headlong: an agent that thinks without pause, where a human message is just one more observation in the stream. Under 10K lines of Bash

Laude's announcement (24.08), **120 points**, thread.

The idea is set directly against the usual reactive harness. Verbatim:

> *"Most agent harnesses are **reactive**... Some harnesses add **cron jobs or heartbeats** that wake the agent on a schedule to run a **fixed checklist** and then put it back to sleep. In Headlong the agent is **never asleep** and there is **no checklist unless the agent creates one**."*

The second sentence describes how most personal bots are built: a scheduler plus fixed jobs. The authors name a third option, **persistent agency**, a continuous stream of thought inspired by inner monologue; **a human message drops into the stream as an observation**, and the agent decides for itself whether and when to answer.

What came of it in practice (their own agent is called Audel and has lived in Slack, Telegram and mobile for a few weeks):

- **One stream of thought for everyone**, no per-user sessions. The agent "connects what different people are working on".
- Unprompted, it **reviewed two unmerged branches of colleagues and found a hardcoded model name** in one of them.
- On its very first day it **audited eight stale branches** belonging to one colleague, and ten minutes later **wrote again to correct its own count**.

- There was a case where a colleague asked it to pass a message to someone else: **the agent refused**, and the status line showed that it had read the request and deliberately decided not to reply.

And the honest half, which they do not hide (the sections are called "What broke" and "Cost"): this is **alpha research software**, it should be run **in a sandbox**, because the agent "can and will run shell commands", with **a separate key on a spend limit**, because it "thinks around the clock". And in plain words: "We **don't share sensitive secrets** with our Headlong agent, and we recommend you don't either".

Why it matters. The architectural fork is named out loud here, and most personal assistants sit on the reactive side deliberately: a session lives while there is a conversation, closes at night, memory is consolidated in a separate run, and proactivity goes through a scheduler. That is cheaper, more predictable and does not burn tokens overnight.

But their argument against the checklist hits a real weakness in that scheme. The scheduler wakes the agent with fixed text, and the wording of the task can diverge from the state of things at the moment it fires. Headlong is not the answer to that (secrets plus continuous token burn), but **they named the problem more precisely than it had been stated before**. Tag [fuzzy]: a research prototype with a few weeks of operation by one team, and no independent measurements.

TOPIC 5

Apple: M6 on 2 nm and the first quad-die M5 Ultra. Large local models got 1.2 TB/s of memory

Apple's press release - **1027 points**, the top tech story of the day (only Dolly Parton's obituary ranked higher). Plus separately Mac Studio (725) and Mac mini (457).

Numbers from the primary source: **M6** is Apple's first chip on **2 nm**, a 12-core CPU, a 12-core GPU with Neural Accelerators, a **dual 16-core Neural Engine**, up to **170 GB/s** of memory. **M5 Ultra** is the first **quad-die** part via UltraFusion, up to **36 CPU cores / 80 GPU cores** and **1.2 TB/s** of unified memory, **50% more than M3 Ultra**.

Why it matters. The **1.2 TB/s** figure is exactly the wall local inference of large models runs into: yesterday's item 6 was about Qwen 27B doing reverse engineering in 30 minutes on **128 GB locally**. For that class of task this is the upgrade. Tag [proven]: manufacturer specifications.

TOPIC 6

ChatGPT learned to sign in to sites by itself, "without seeing" the login and password. Plus triggered tasks and a $100 business seat

Three OpenAI announcements in one day, and the first is about the limit of trust in agents.

ChatGPT Work now signs in (25.08, 21:40, **780K views**), confirmed by Brockman: the agent uses a computer and a browser to sign in to sites **on web and mobile**, and does it "**without ChatGPT ever seeing your username or password**". Their own examples: setting up utilities in a new apartment, booking a DMV or passport appointment.

Second, triggered tasks (26.08, 01:55): scheduled tasks can now **react to a change in Slack, Gmail and GitHub**, on top of running on a schedule; free users got scheduling too.

Third, ChatGPT Business Premium at $100 (1.1M views): no five-hour limit. This continues yesterday's story about the return of 5-hour Codex limits for Plus.

Why it matters. Triggered tasks have long been built in-house by anyone with a conditional scheduler: the gate fires, the agent is woken; it does not fire, silence. There is no gap here. Signing in without showing the password is a separate trust architecture, and few will copy it: the rule "secrets do not go into the chat" is simpler and can be checked by eye, while auto-login on the user's behalf cannot. Worth noting that the industry is moving that way, and that diverging from it can be a deliberate choice. Tag [proven] on the fact of the announcement; how it actually works and what the sandbox sees is **unverified, the product is not in hand**.

TOPIC 7

C2PA on Android is broken in a way that cannot be fixed. AI slop signed by a "real camera", following a blog post

David Buchanan (retr0id) (25.08), **106 points**. C2PA is the standard where "a camera cryptographically signs the shot", sold as the cure for AI fakes.

The chain of reasoning is short and hard: C2PA apps on Android rely on **Key Attestation / Play Integrity**, root exploits break that model, **root is achieved through cheap hardware fault-injection attacks**, and hardware holes in existing devices **do not get patched**. The author's conclusion: "C2PA on the Android platform is **broken, in a way that cannot be realistically patched**".

And why Android specifically: Google themselves are quoted, Pixel Camera has **Assurance Level 2**, the highest in the C2PA conformance program, and on mobile it is **only possible on Android**. So the strongest implementation is the one under attack.

The sharpest part: "Thanks in part to **LLMs**, root LPEs are coming out **faster than Google can ship patches**". At the time of writing there is a **one-click root for fully patched Pixels** (CVE-2026-43499). The article contains an AI-generated image that C2PA **certifies as a genuine unretouched photo from Pixel Camera**, and a YouTube video with an infobox reading "captured with a camera". The detail that stands out most: **at 19:12 the same day Google manually removed the "Captured with a camera" badge** from that video, and the author added this straight into the article.

Not a 0day: reported at least 90 days ago.

Why it matters. A direct line to **yesterday's item 5** (MS Paint embedding a server-side GUID in local images) and to **21.08** ("AI blindness"). Three days, three stories, all about how **a technical provenance mark does not work as proof**: in one case it is applied without the user knowing, in another it can be forged by following instructions. The practical conclusion is the same one already applied to links: **proof moves into the content**, the badge guarantees nothing. Tag [proven]: with reproduction instructions and a public CVE.

TOPIC 8

Stanford updated "Canaries in the Coal Mine": employment among 22-25 year olds in AI-exposed occupations is 19% lower. A year ago it was 13%

Ars Technica, **137 points**. An update to "Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence", August 2026.

The numbers:

- Employment among **22-25 year olds** in the most "AI-exposed" occupations is **19% lower** than among peers in less exposed ones. Last year the gap was **13%**.
- Since 2022, in the top 40% of "AI-hit" occupations youth employment **fell by ~11%**, while in the remaining 60% it **rose by 10%**.
- **Across the economy as a whole there is almost no difference**: the effect concentrates on the entry point into the profession.
- The mechanism is **non-hiring**: the pace of hiring newcomers drops, with layoffs unchanged. And it hits **the number of jobs**, while wages do not move.

The data is anonymised ADP payroll; "exposure" was computed partly through the **Anthropic Economic Index**.

Why it matters. This is the strongest evidence on the topic Faye raised yesterday ("coding expertise is going to collapse", UPenn, 1000 students, −17%), and today's figure **joins the other end of the same arc**: yesterday was the mechanism (without friction expertise does not grow), today it is the consequence in the labour market (the entry point narrows). For a product this is the applied question of hiring juniors, and it now has a number. Tag [promising]: this is a preprint update with revised statistics, not a peer-reviewed publication; and correlation with "exposure" does not directly prove causation, as the authors themselves are careful to note.

TOPIC 9

"How much of HN is AI": ~50% of top stories every day. But the article itself is from March, and you only see that if you open it

lcamtuf, **253 points**, the third tech story of the day by points.

The measurements: in February 2026, **40%** of the daily HN selection was about AI or written by AI; in June it was **~60%** early in the month and **~50%** by the end. Generated text was caught with **Pangram**; everything flagged was checked, and the detector turned out to **miss** more often than over-flag.

And here is the main thing missing from the HN headline: the article is dated 12 March 2026, and the freshest data in it is from June. What sits on HN today is **an old text that resurfaced**. That only became visible after downloading and reading it; the headline shows nothing.

Why it matters. First, this is a direct comment on the digest's main source: if half the HN top is about AI, then the "points>80 + is this actually useful" filter is doing exactly the job it exists for. Second, it is **a cheap lesson about deduplicating by date**: the window is counted by the submission's `created_at`, and today's case shows a difference of **five months**. It is placed ninth, with the caveat in the headline. Tag [proven] on the method, [fuzzy] on how current the numbers are: June was three months ago.

TOPIC 10

Qwen 3.8-Flash-Next ships "tomorrow": 312 points for an announcement that does not exist yet. The parameters in the HN headline are unconfirmed

The ModelScope page, **312 points**, thread. Clement Delangue reposted it with "Who's excited?", and another post counts "22 hours to go".

Checking what the HN headline claims ("125B a6B") found no confirmation. The ModelScope page returns an empty JS shell without curl; the HuggingFace page exists (200 against 401 for a deliberately invented one, so the check here measures something real), but **the model API is closed**. And the most telling part: **the first comment in the thread** asks exactly the same question, "Can you share the source for the parameter count (125B A6B)? I didn't see it anywhere in the page". The figure in the HN headline came **from the submitter**, and 312 points do not make it true.

Why it matters. It is placed tenth for exactly that reason: the topic is real (the model is announced, the pages are live), the specifics are not. This is the same class as "HN points ≠ proof" from 24.08, only in reverse: the points are there, and there is nothing to verify. Tag [fuzzy]. It comes back when the weights ship.

misc: Garry Tan's jump on memory across harnesses - "cross-harness memory maxxing", a person putting `gbrain` into Claude Code / Codex / grok build at once · A post on the gap between models and workflows (26.08, 04:17) · str.lower as a security vulnerability in Python by @sethmlarson, 89 points - `stringprep` has to stay on Unicode 3.2.0, while `.lower` follows the interpreter version · A month of Grok Bot given away to annual subscribers (314K views) · Anthropic asked its SF office to work from home over a possible security staff strike, 119 points · Firefox 157 enables JPEG XL on all platforms, 313 points · The Grok Bot

source + agentmail + Stripe Link as a "full personal agent" - interesting, but it is a thread advertising three products at once

---