

GitHub published the postmortem of the 17.08 outage: the cause was volume. Commits went from 1.4B to 2.9B a month since April

AI digest, Friday 21.08 - three stories of the week closed with numbers: GitHub named the real cause of its outage, Stripe disclosed the OpenRouter price, and Asana showed what a migration by agents costs.

Friday, 21 August 2026

IN THIS ISSUE

1. GitHub published the postmortem of the 17.08 outage: the cause was volume. Commits went from 1.4B to 2.9B a month since April
2. The OpenRouter price is finally named: over \$8B, mostly in stock. And in the same letter Stripe tells investors that "the singularity has started" - on 1 January
3. Asana tore out a test framework in 2 weeks instead of 5 years - for \$12 thousand instead of \$6 million. And the same day Orosz independently confirmed it: Airbnb and Uber see the same
4. A poisoned Rust crate with 245M downloads: the attack is built so that Cargo's own warning pushes you to the infected version
5. "Please don't paste AI at me" - 997 points and 549 comments in a day. The most popular page of the day turned out to be one page of text about manners
6. Bun 1.4: rewritten from Zig to Rust, and there is a figure about Claude Code specifically - CPU usage there dropped by half
7. Slack launched Slack Code: agents from Anthropic, GitHub, Cognition and Vercel sit in a channel alongside people
8. Mollick on the sameness of machine text: "even when LLMs write well, the lack of style diversity is devastating"
9. Karpathy on why abstractions are worth tearing out: "PyTorch is a so-so intermediate language"
10. Google is buying a bankrupt airline's data - a continuation of yesterday, and now the genre is visible

TOPIC 1

GitHub published the postmortem of the 17.08 outage: the cause was volume. Commits went from 1.4B to 2.9B a month since April

396 points on HN, [the official writeup by GitHub CTO Vladimir Fedorov](#). [HN thread](#).

This continues item 1 from 18.08. Back then the event came off the status page with a note that there was no RCA yet. Today there is one, and it changes the story.

What it turned out to be: • Duration – **7h 47m**. On 18.08 the status page said "7h 35m"; the official writeup gives 12 minutes more. The figure now comes from the primary source.

- **"Neither outage was caused by a code or configuration change"** – verbatim. Both August outages (6.08 and 17.08) were **capacity failures**: a critical component in Central US did not scale to the new traffic peak.
- **Commits: 1.4B → 2.9B a month since April**. Fedorov adds right away: "That growth explains the pressure on our systems, but it does not excuse these outages".
- What they have already thrown at it: **3M+ CPU cores, 120 petabytes** of fast storage. Azure now carries **58% of the platform's load**, against 12% in May.
- Copilot took longest to come back because of a **client-side retry loop** that raised traffic during recovery. They are now introducing shared limits and retry budgets across services.

Why it matters. The valuable part is 1.4 → 2.9B. That is a measured trace of agents inside someone else's infrastructure: code volume doubled in four months, and **capacity** broke first. Yesterday's Linear item (agents write half the tasks) and this one are one story from two sides: there you see how much agents do, here how much it weighs in hardware.

The second, smaller lesson: **a client retry with no budget turns degradation into an outage**. Any scheduler that hits three or four external APIs on every tick has retries and rarely has a limit on them. Same pattern.

TOPIC 2

The OpenRouter price is finally named: over \$8B, mostly in stock. And in the same letter Stripe tells investors that "the singularity has started" – on 1 January

Axios, with the investor letter in hand – Dan Primack and Ryan Lawler. The reaction from **@emollick** is the best of the day.

Three days of dedup in a row, and this is the finale: • **18.08** – the deal ran as item 3 marked "Bloomberg, paywall", **with no link**: a check showed bloomberg.com returns 403 even for a deliberately made-up URL. The figure circulated as "\$7B, unconfirmed".

- **20.08** – OpenRouter confirmed the deal on its own blog but **did not name the price**.
- **today** – Axios has the letter itself: **over \$8B, mostly in stock**. Waiting for the signature instead of running an uncheckable link turned out right three times running.

The rest of the figures from the letter (primary source, not a retelling): • first-half revenue **+41% year over year**, free cash flow **+43%** • **88% of the Forbes AI 50** – OpenAI and Anthropic included – build on Stripe • the share of revenue from AI and crypto companies **more than doubled** in a year • the share price compounds at **31% a year** since Series D, against 14% for

the S&P 500 and 18% for the Nasdaq • IPO - **on an indefinite pause**: Stripe writes to investors plainly that staying private at a moment like this pays better

And now the antidote, because this is a company letter about itself. Stripe told investors that 1 January 2026 marked "the beginning of the singularity". Mollick hits the seam exactly:

"Either the Singularity is now a marketing term, or Stripe believes we are at the point 'in the history of the race beyond which human affairs, as we know them, could not continue'? That somewhat makes forward-looking financial statements meaningless."

(In brackets he notes he is quoting **von Neumann's original wording** - the man who coined the term.) The sting is that a company cannot simultaneously promise investors predictable financial forecasts **and** announce an event after which, by definition, nothing is predictable.

Why it matters. Two practical things. First, **OpenRouter is now part of Stripe and costs \$8B**; the "nothing about your integration changes" promise from yesterday's blog now has a price tag, and after a sum like that a product usually starts drifting toward whoever paid. Second, Stripe's fresh **Provisioning API**: agents will be able to **open accounts themselves** at Supabase, OpenRouter and Cloudflare through code, without a browser. That is exactly what is still done by hand through Playwright. Early access by email for now.

TOPIC 3

Asana tore out a test framework in 2 weeks instead of 5 years - for \$12 thousand instead of \$6 million. And the same day Orosz independently confirmed it: Airbnb and Uber see the same

The OpenAI case study. @gdb, 109K views. Independent confirmation - **The Pulse by Gergely Orosz.**

Figures from the case: • removing **Enzyme** (a dead test framework that blocked frontend upgrades) - the plan was **at least 5 years** and **~\$6M** in headcount • done in **1.5 weeks of actual effort**, spread over 2 calendar weeks • model plus infrastructure cost **~\$12K**. That is **500 times** cheaper than the staffing-plan estimate • the setup: **a five-sentence prompt** → up to **four agents in parallel**, each in **its own copy of the repository**; an engineer checked progress **twice a day** and reviewed **every change** • one line stands on its own in the case: **"Simpler instructions worked better than a more elaborate setup"**

Asana CTO Amritanshu Raghav sets the limit himself, which is honest for a vendor case:

"Not every multi-year project will collapse into weeks. But agents give engineers more room for craft, and make once-impossible work worth attempting."

Why this case is not filed as a press release. It is an OpenAI case study about an OpenAI customer, an advertising genre by definition. So the search went further, to someone writing about it from outside the seller's side. Orosz covers this exact story the same day and adds that **Airbnb and Uber report similar results**; his conclusion is that AI looks like a

"superb fit" for **framework migrations** specifically. The body of the issue is paywalled, the free part carries this point in full.

Why it matters, and where the limit is. The task class here is narrow and recognisable: **mechanical, bulk, with a clear done criterion** (the tests are green or they are not). "Rewrite 5000 identical spots." That debt sits in any codebase that grew by accretion.

But two limits from the case itself are easy to miss behind the \$12K: **the engineer reviewed every change**, and **the parallel agents sat in separate copies of the repo** - exactly what broke the swarms in Anthropic's 17.08 research (18 of 30 agents created the same branch). What works is **agents plus isolation plus human review at every step**.

TOPIC 4

A poisoned Rust crate with 245M downloads: the attack is built so that Cargo's own warning pushes you to the infected version

421 points on HN, [SafeDep's technical writeup](#), [thread](#).

The mechanics, and they are unpleasantly clever: • on 20.08 a compromised release of `arrayref 0.3.10` appeared on crates.io after the maintainer's account was taken over • it pulls the typosquat `proc-macro1` (not `proc-macro2`, the real one), whose build script **downloads and runs a binary right during compilation**. Simply **building** the project is enough • the fake crate's `src/` is a **genuine copy of proc-macro2**, so builds kept working as if nothing had happened • the author is forged as **David Tolnay** (`dtolney` against the real `dtolnay`), with his name written into the metadata • **the nastiest part:** the attacker **yanked the clean releases 0.3.5 - 0.3.9**. For a yanked crate Cargo prints "consider updating to a version that is not yanked", and the only unyanked version left was the **infected 0.3.10**. The package manager's standard warning worked as the delivery mechanism • **scale: 245M downloads** all time; the crate sits transitively under `tiny-skia` and `winit`, which means under most GUIs on `egui / eframe / iced`

Why it matters. This is the third case in a week from one storyline: 18.08 a Copilot agent removed an injection guard, 19.08 (via DOU) LiteLLM and 195TB of secrets from a poisoned PyPI, today crates.io. What they share is that **execution happens at build time**, and no code review catches that. The practical conclusion goes wider than Rust: a pile of scripts with a `requirements.txt` and no hash-pinned lock file is exactly the class where a swapped version slips through quietly.

TOPIC 5

"Please don't paste AI at me" - 997 points and 549 comments in a day. The most popular page of the day turned out to be one page of text about manners

[dontpastetheai.com](#), [HN thread](#) - the second highest scoring item of the day.

A site in the spirit of `nohello.net` : a link you send to someone who just dumped a wall of chatbot text on you. The claim, verbatim:

"When someone asks about something, they want **your** answer. Not a wall of ChatGPT text. A short answer from a human is always worth more than an answer from a robot, even when the robot is right."

And the sharpest bit: "**the person on the other side has the same tools**. If they wanted a generic answer they would have had it in a couple of seconds. The question was asked because they wanted a **personal** view: context, taste, judgment".

What it suggests instead: draft with AI, **but read it and rewrite it in your own words**; take only the part that actually answers the question; if a piece from the model is good, **quote it and explain why**. And if you have nothing to add, say so. Silence is an option too.

Why it matters. On 20.08 item 5 carried the bug report "Opus 5.0 sends incoherence into the stratosphere" with a linter against "load-bearing" in the build, a complaint about **machine language**. This one is a complaint about **machine mediation**: a person turning into a pipe between the model and whoever they are talking to. The second is worse because it is invisible.

The **fresh tooling news** belongs here too: Claude Code v2.1.237 (20.08) added a built-in output style, "**Concise**" - "leads with the result, skips preamble and narrativization, does the work just as thoroughly". The complaint from yesterday's item already has a switch in `/config`. Checked against the **releases**, not rumour.

TOPIC 6

Bun 1.4: rewritten from Zig to Rust, and there is a figure about Claude Code specifically - CPU usage there dropped by half

236 points, **the official release**, **thread**.

- **+1517 tests** from the Node.js suite, the biggest compatibility jump since Bun 1.0; over **2900** issues fixed
- idle CPU **5 times** lower, memory **down to -35%**, startup on Linux **50% faster**

verbatim from the release: "For **Claude Code**, a large long-running application built on Bun, production CPU usage dropped by 2×: p99 from 24% to 10%, p50 from 5.8% to 2.5%"

- new: `Bun.cron`, `Bun.Image`, `Bun.WebView`, `Bun.markdown`, `bun audit fix`, `bun run --parallel`
- `node:events`, `node:sqlite`, `node:trace_events` - **100%** of Node tests;
- `node:http` / `fs` / `stream` - 97%

Why it matters. Two lines deserve attention. First: **Claude Code runs on Bun**, and p99 CPU falling from 24% to 10% is physically felt wherever sessions hang around for hours. Second: **Bun.cron** - a scheduler now lives in the runtime itself, which makes a separate layer for it optional.

TOPIC 7

Slack launched Slack Code: agents from Anthropic, GitHub, Cognition and Vercel sit in a channel alongside people

Salesforce's [official page](#), [The Verge's writeup](#), Benioff's announcement. The best comment is [@levie](#), and it is about the substance of the form factor:

*"One of the most interesting dynamics with AI agents is finding the right form factor for **how exactly they should show up inside the workflow**. The longer the tasks agents can carry and the fewer interventions they need, the more they can live in the tools people already use."*

Why it matters. Homemade wrappers that spawn an agent from a chat have been around for a while: the form factor was guessed before Salesforce got there. The difference is that the vendors give you a **channel**, where an agent and people in one thread see the same thing, while the homemade version is usually a **private chat with a single session**. "Several people and several agents in one working thread" is precisely what the product closes.

The other side is competitive: once agents in Slack become the standard, your own integration stops being an advantage and becomes hygiene. What was built by hand has just turned into a product at three large vendors at once.

TOPIC 8

Mollick on the sameness of machine text: "even when LLMs write well, the lack of style diversity is devastating"

[@emollick](#), 300 reposts, 14K views - plus a [follow-up](#) on why settings do not fix this.

"Reading the same prose in instructions, and on social media, and in ads, and in software, and in presentations, gets nauseating over time. Prompting only takes you so far. We need real diversity (and it is understudied)."

The key is in the second post: **changing temperature and top-p is not enough**, the sameness comes through anyway. Then a claim wider than style: diversity is needed if you want AI to **produce different ideas, do science** and so on. Uniform output is a **search** problem: a model with a narrow distribution will not find what nobody prompted for.

He also [recommends separately](#) replacing the default writing and design skills in harnesses with your own, "for your context and taste". That continues his own post from yesterday about the skill creator.

Why it matters. Three items today are about one thing: N^o5 (do not be a pipe), this one (the model sounds the same), and yesterday's linter against "load-bearing". The working antidote is simple too: instead of a default writing instruction, a voice profile assembled from the author's own texts, with measured frequencies and a rule against changing tone.

TOPIC 9

Karpathy on why abstractions are worth tearing out: "PyTorch is a so-so intermediate language"

@karpathy, 200 reposts. The claim:

"It is increasingly attractive to tear out these abstractions now that agents can take on a lot of the math, the grind and the checking. Many abstractions were built for a world constrained by finite intelligence and attention in the field."

And further: the extrapolation is that the spec becomes something like `microgpt`, scalar Python with loops, and the rest is compilation.

Why it stays in the issue. The frame applies well beyond ML: abstractions get built because a person cannot hold the details. Once an agent takes part of the attention, some layers become technical debt. The reverse is visible too: a single dispatcher over a dozen scattered scripts is justified precisely because the contract is visible from `help`, while the scripts underneath it are a layer left over from when the interface was remembering paths.

TOPIC 10

Google is buying a bankrupt airline's data - a continuation of yesterday, and now the genre is visible

Yesterday this was item 5 (\$10M for 100M emails and 30M call records). Today the same storyline came in sideways: **Stripe writes that 88% of the Forbes AI 50 run on it**, Google buys up archives, Nvidia puts the first Vera Rubin racks into OpenAI (@gdb, 70K views).

Separately, as a **raw claim for tomorrow**: OpenAI launched the **AI Futures** blog, a Strategic Futures team writing about "risks of concentration of power". The first post, by Dean Ball: the state has historically rested on the consent of people, because both the army and taxes leaned on human labour; if both get automated, "the people may be left without a seat at the bargaining table". The piece was read in full. **It is not filed as its own item**, because the text carries an explicit disclaimer: these are the author's views, an essay, and OpenAI does not call it its position. But the genre of "a lab opens a policy department" is worth noting.

misc: **Aaron Swartz against Meta: 1205 points** - prosecution for scraping against unpunished scraping for AI · **AliExpress quietly fingerprints via WebAudio** and breaks Bluetooth multipoint in headphones, 927 points · **@levelsio measured himself**: with no girlfriend, recovery -17%, sleep -4%, productivity +31% (n=1, but the data comes from WHOOP and home sensors) · **Sid Sijbrandij, GitLab founder, on his own cancer**: "n=1, not a model to copy, but proof that fully personalised medicine is viable" · **Ro Khanna: 5402 stock trades worth \$165M** - bought Moderna on 30.06, two weeks before +173% on the vaccine news · **Claude Code v2.1.238**: `keybindingFlavor: readline` for Ctrl+W as in Bash · **tetrachromacy** - some women have four cone types instead of three, and usually do not know it

