AI DIGEST

"Everything I own, owned": five devices taken apart by an agent on Opus 5 in 13 hours. 1368 points, the top story of the day

AI digest, Tuesday 25.08 - the day agentic reverse engineering stopped being a demo: the top HN story was done by the same model in 13 hours, and yesterday's market puzzle got a concrete answer, which has nothing to do with model quality.

Tuesday, 25 August 2026

IN THIS ISSUE

TOPIC 1

"Everything I own, owned": five devices taken apart by an agent on Opus 5 in 13 hours. 1368 points, the top story of the day

Article by schlarpc (23.08) - **1368 points on HN**, thread.

Over two weeks of evenings the author fed an agent the firmware of five peripherals lying around the desk. The model is named outright: **Claude Opus 5**. And, rarely, the effort is measured from Claude Code session transcripts:

DEVICE	AGENT TIME	HUMAN PROMPTS
Shure MV7 (microphone)	4.2 h	32
Insta360 Link (webcam)	3.7 h	33
Elgato Key Light Mini	2.4 h	10
Elgato Cam Link 4K	1.5 h	10
ASUS ROG Swift PG42UQ	1.2 h	13
Total	**~13 h**	**98**

What turned up (most of it verified live on the hardware):

- **The webcam records with the LED off.** The firmware holds a table of LED "patterns"; the agent wrote a tool that patches the entry for the "recording" state, recomputes the MD5 and flashes it back. There is no tamper protection, just an appended MD5. On top of that, a USB vendor class channel gives arbitrary file read/write and reboot, so **flashing takes no user action at all**.
- **The microphone carries a full plaintext command shell** - 48 commands over USB HID, reachable even **from a web page via WebHID**. There is a four-level privilege system, and the entire "authentication" is a string comparison against the level name: literally "`su` `sup` **just works**". The top level can disable the mute button and **drive the mute indicator independently of whether the microphone is actually muted**.
- **The Cam Link 4K the author ran fully unattended**: started it before bed, woke up to a finished teardown and a working firmware updater.

Why peripherals turned out to be the perfect target: "they're **tiny computers attached to my computer**, with a data connection to the host and usually a firmware update mechanism, so **an agent has something to iterate against**".

Why it matters. First, **this is a direct productivity measurement on this stack** - a real task with validation on hardware: 2.6 hours and ~20 prompts per device on average. Next time the question is "how long will this take", here is a calibration point.

Second, it is **a change in the threat model for home hardware**. The key comment in the thread, @teddyh, quotes the author: WebUSB/WebHID/WebBluetooth mean that "**a moment of user indiscretion in accepting a permissions prompt could permanently backdoor one of their attached devices**". That used to cost a reverse engineer weeks; now it is an evening. Practically: the recording light on a webcam is no longer proof that the camera is not recording. Tagged [proven] - the code is in public repos and was tested on live devices.

TOPIC 2

Yesterday's item 2 puzzle is solved: it is ZDR. "Procurement picks the model, not eval scores"

A bridge back to yesterday's issue, which carried the Ramp data - Opus 5 at 3.5% of spend against 28% for Opus 4.8 - with an honest admission that the cause was unknown. Within a day it turned up.

Martin Casado (a16z) asked out loud (24.08, 07:39 - **188k views**): "Lots of folks jumping to conclusions on the FT market share data for Fable. **The reason is not at all obvious to me.** Is it due to cost? Speed? Aggressive moves from OAI? ZDR? Open Source? Bad press / vibes? Refusal to work in some domains? Biased data?"

Seven hours later, his own conclusion (15:08): "**Consensus on this seems to be ZDR**, which tracks to my own experience as well. If that really turns out to be the case, it's an **incredibly strong and clear message to the labs** on the impact of data retention policies".

ZDR = zero data retention, the contract clause "we do not store your data". Aaron Levie spells out the mechanics (24.08, 15:23, **76k views**): "It's not fully appreciated that **ZDR is responsible for a substantial amount of the growth of AI**, because it dramatically simplified the compliance process most companies deal with in handling data with subprocessors". Then the specifics: applied AI tools agree with customers to offer **ZDR models only**, because otherwise every model has to go through its own exception; and inside companies the rules often say "ZDR only" outright, because **separating PII from the rest of the context window is technically beyond them**. Levie's conclusion: "**Without ZDR, AI diffusion grinds to a halt**".

The most concrete piece comes from a procurement practitioner's reply, @iabdul_wasey: "ZDR is the first line of every security review, and a model without it turns a **two week approval into a quarter**. Teams pick the **second best model with ZDR** over the best one without. **Procurement picks the model, not eval scores**".

What stays unverified, stated plainly: this is **industry consensus on Twitter**. There is no document. Casado himself hedges ("if that really turns out to be the case"). Anthropic has not confirmed it officially, and the FT article itself is still unreadable (ft.com returns 403 for both the real URL and a deliberately invented one, checked yesterday). Hence [promising], not [proven]. But the mechanism at least **explains the anomaly**, which none of yesterday's hypotheses did.

Why it matters. ZDR is a corporate story, so there is nothing to apply today. Two takeaways are worth keeping in mind though. First: **the best model loses to the faster legal sign-off**, so when picking a stack for a customer-facing product, the retention question comes before the benchmark question. Second, broader: yesterday no explanation was invented when there was none, and within a day one arrived from people in the market. Same principle as with other people's numbers: **an honest "I don't know" beats a plausible invention**.

TOPIC 3

a16z on OpenRouter data: agents burn 5x more tokens than people, and 85% of that is cache

Charts of the Week from a16z - reposted by Greg Brockman (24.08, **146k views**) with the caption "a new way to get knowledge work done". The tweet only had the per-profession multipliers; the interesting part sits deeper in the report.

The main numbers (sources named: OpenAI, OpenRouter, Similarweb data):

- **Agents use nearly five times more tokens than people**, and agentic usage has grown **~14x since February**.
- **Over 85% of agentic token burn is cached prompt**, and cached tokens account for almost all of the relative growth. A person works in prompt-and-answer mode; an agent **iterates toward a goal**: context is loaded once and then read and written incrementally.
- The gap between a typical and a top company is **~8x in token output**; in tech, the top decile outputs **32.5 times more tokens** than a year ago.
- Codex adoption by profession since February 2026: **lawyers 108x**, sales 41x, recruiting 41x, marketing 26x, medicine 24x. The fastest growth is **not among engineers**.
- A side effect few noticed: **traffic to legacy automation is falling**, with N8N, Zapier and Make all showing double-digit declines over the last 12 weeks.

Why it matters. The 85% cache figure speaks directly to the economics of the design and explains why the "heavy stable context plus many small steps" pattern is cheap: **a large unchanging prefix (instructions, reference material, memory) gets cached, and you mostly pay for the increment**. It is an argument **in favour** of big stable instruction sets, and it gives yesterday's Breunig claim about the harness mattering more than the model a financial floor as well.

Second: the Zapier/n8n decline is an early illustration of yesterday's Garry Tan forecast. The post-Zapier world is already here - scripts, schedules and instruction sets instead of visual workflows - and it arrived because it turned out cheaper and more flexible.

TOPIC 4

"Coding expertise is going to collapse": 501 points and three studies, one of which measures the harm at 17%

Essay by Lars Faye (24.08) - **501 points**, thread. The text rests on three cited studies.

The central paradox is put cleanly: "If these tools **demand expertise**, yet the tools can actively **circumvent the friction that cultivates expertise**, then what is the path for one to become an expert so they can effectively use these tools?"

The studies:

- **JetBrains cites "The Widening Gap"** (analysis of live beginner sessions): "Participants thought it was like having a personal tutor. From the data in our study... we observed that

they did not, in fact, use GenAI tools like a personal tutor. In fact, it was quite the opposite". Those who leaned on AI harder "often **skipped crucial planning stages**" and ended up with "**illusion of competence** rather than true understanding".

- Those who coped had what the authors call "**negative expertise**" - "the ability to **ignore incorrect or unhelpful GenAI suggestions**".
- **UPenn, 2025, "Generative AI without guardrails can harm learning"**: 1,000 students, and those who studied maths with an unrestricted LLM scored **17% worse** than those who had only a textbook, while **believing they were learning better**. In the same study the "GPT Tutor" version (the model helps, the student solves it) produced the opposite result.

The soberest comment in the thread came from the author, answering "so how do you teach then": he points to **Uncle Bob**, who is "as AI pilled as they come" but says juniors **should not touch AI tools for the first three years**.

Why it matters. For an experienced developer this is the "ossified knowledge" case, where the tool is a clean win. Two other things are interesting. First, hiring: "can work with an agent" and "understands what the agent did" are different criteria, and the second one does not show up in a demo. Second, teaching children: the study's conclusion is specific, **the "give me the answer" mode hurts, the "help me, I solve it" mode helps**. When a student asks something about their course, the right behaviour is Socratic, without the finished answer. Tagged [proven] for the studies themselves, [promising] for the author's extrapolation to the industry.

TOPIC 5

MS Paint embeds an invisible watermark with a server-issued GUID even in locally generated images. 613 points

Reverse engineering by Xusheng Li (24.08) - **613 points**, thread.

The mechanics, pulled apart with a disassembler: Paint and Photos **really do carry local models** (four `.onnxe` files, the largest 302 MB - XORed ONNX, with the key sitting in `segapi.dll`). But before local generation `AIServices.dll` **sends the prompt to a Microsoft server** for moderation, and the server returns a **GUID**. That GUID (16 bytes) is then written into the pixels themselves via `Watermarker.dll!WmkWriteWatermark`, a content-adaptive block watermark in the SVD style. On a 512x512 test image, **193,376 of 262,144 pixels** changed.

Two details that make this more than a trifle:

- **The separate "visible watermark" setting does not control it.** The visible Copilot logo can be turned off; the invisible GUID cannot.
- If the watermark **cannot be written, Paint turns the whole generation into an error** instead of handing over the image without it. So it is a hard requirement.

Honesty about the source, because it is the first comment in the thread. The submitter (not the author) posted a warning himself: "**AI-generated text warning**", and the second commenter recognises "Claude flavored" paragraphs in the text. The check went on substance. Style was set aside: **the findings themselves are concrete and reproducible** - DLL names, function signatures (`Paint::AI::AddWatermark(Gdiplus::Bitmap&, winrt::guid const&)`), model sizes, error codes `-5` / `-6`, the payload structure `0x4c || GUID || checksum`. That is not the stuff of hallucination. The prose style is AI-ish; the content is a reverse engineer's work. The domain passes the check honestly (**404 on an invented URL**).

Separately: the author writes that he was inspired by a "recent Claude Code text-watermark announcement". **The check does not confirm this**: in the claude-code CHANGELOG.md (fetched as raw text through `api.github.com`, 543 KB, verified to really be `# Changelog`, latest version 2.1.241) the word "watermark" returns **zero hits**. So the Claude Code remark stays unverified, which does not touch the Paint finding itself.

Why it matters. "Generated locally" does not equal "nothing left the machine": the prompt goes to the server anyway. The conclusion is broader than Paint and applies to any "local AI" in an OS - **check the traffic**. The marketing promise counts for nothing here. It is the same class of error that shows up in this digest every day: **an indicator in settings does not equal the real behaviour of the code**.

TOPIC 6

Narayanan did napkin math on data centre bans: a one-year state moratorium = 5-10 hours of delay to AI progress

Post by Arvind Narayanan (24.08, 15:11, **66k views**), backed by Mollick.

The logic: inference efficiency at a given capability level rises steadily, so "N GW blocked" can be converted into the **time** in which the same capability is caught up by efficiency alone. The result: "if a typical U.S. state enacts a 1-year moratorium, it **slows AI efficiency progress by 5-10 hours**". And given that sites are interchangeable (assuming 90% "leakage" to other states), even a full ban in a state like New York is "**less than a day**".

His own follow-up: efficiency delivers **an order of magnitude more progress than data centre growth** (5x/year conservatively against 25%/year), and bans "do nothing about all the capacity already built".

The most valuable part here is the author's footnote. Verbatim: "While I may have gotten some numbers slightly wrong (**I used AI for the analysis and did some spot checks**), the rough order-of-magnitude is enough to make my larger point". He publishes a calculation done by AI **and says so out loud in the same post**. Mollick sums it up: "My AI broadly agrees with Arvind's AI".

Why it matters. This is exactly the discipline this file holds to: **say where a number came from and where its reliability ends**, and do not present an estimate as a measurement. Narayanan does not hide that an agent did the arithmetic, and does not pass an order of

magnitude off as precision. This is a model for how to phrase things. It is not a fact about US policy.

TOPIC 7

Mollick: "yes, they can tell" - a reviewer spots AI because the ideas start to rhyme

Post by @emollick (24.08, 20:54, **31k views**, 425 likes): "Note to anyone sending in an application, paper, or any other document where **one person reads many submissions: yes, they can tell**. It is obvious in most cases that you are using AI, **even if the language is changed**, as the ideas & concepts start to **rhyme as you read lots of them**".

And he immediately qualifies himself: "I don't think there is anything wrong with getting AI help in writing or feedback or whatever... but **if you use it to do too much, it starts to obscure your unique contribution**".

The mechanism is in the **distribution**: one application looks fine, but across a hundred they share the same argumentative skeleton. The smartest reply in the thread, @zazmic_inc: detection works better **by comparing texts against each other** than by passing a verdict on a single text.

Why it matters. For any public writing the risk is one and the same: **losing the thing people read the text for**, the concrete numbers from practice and the author's own phrasing. This backs the "don't change the tone" rule with a technical argument: the tone is the part that does not rhyme with the rest of the feed. Drafting, yes; rewriting in the model's voice, no.

TOPIC 8

Levie and Tan close yesterday's "systems of record to harnesses" thesis from the data side

A continuation of yesterday's item 9. Yesterday Garry Tan tossed out a forecast in one sentence (a day on: **384k views**, 2.6k likes, three times the earlier figure). Today Aaron Levie replied (25.08, 03:12 - fresh, already 20k views):

"Systems of record have never been more important than in a world where you have AI agents that will do **100X more work** on these platforms than people ever did. Agents will be querying the data in these systems, processing tasks, executing workflows, collaborating with human and..."

And Tan himself sharpened his thesis: "There will still be **API's and acl's and sql and underlying data structures that are deterministic**. It's just that software companies have to build the AI harness and full solution for their customers **or be subsumed by it**".

So in a day the forecast went from "systems of record will die" to something more precise: **the data and the deterministic access are going nowhere; what dies is the monopoly on**

the interface to them.

Why it matters. Tan's clarification describes the architecture people already arrive at in practice: **a SQLite database, spreadsheets, device APIs are deterministic systems of record, and the set of commands and instructions on top of them is the harness**. The table is not replaced by the agent; the agent is given disciplined doors. Item 3 (85% cache) explains why this is also cheap.

TOPIC 9

Xiaomi Xring O3 has caught up with Apple cores: 44 MB of cache, 21 execution ports. 777 points

Teardown by Daniel Lemire (24.08) - **777 points on HN**, thread. Xiaomi's new chip "roughly matches Apple cores on single threaded tasks, and is **much faster in multithreaded** execution".

The specifics: **44 MB of cache in total**, more than most laptop processors; the big C1-Ultra cores support **SME2** (matrix extensions for AI) and SVE2; the core is "astonishingly wide" at **21 execution ports**. Lemire adds a caveat right away: Apple is about to announce its next chip, so the lead may be short, and finding a phone with an O3 will be hard.

Why it matters. Nothing to apply, it is background to item 3 from yesterday (local models): **SME2 in a mobile chip means on-device inference is becoming the norm**. The HN thread is mostly about geopolitics; that part is not covered here, it is outside the digest's scope.

TOPIC 10

Small but real: IPFS lost its funding, and Replit Agent versus Claude CoWork

IPFS. Shipyard announced it is winding down (24.08) - **339 points**, thread. Protocol Labs is not renewing the funding; the last day of work on IPFS is **30 September 2026**. Their wrap-up carries numbers showing the team was alive: rebuilding the gateway infrastructure handled **~3x more traffic at ~80% lower cost**. The thread went almost entirely into arguing that "an NFT is just a link". The technology itself was barely discussed, which is an epitaph in its own right.

Replit. Amjad Masad claims (24.08, 16k views): "Replit Agent completely **replaced Claude CoWork** in my day-to-day work. It's much more persistent and fastidious, and it **uses code/software more effectively** to complete tasks". This is a CEO on his own product, a marketing line delivered as a position. The interesting part is the criterion of comparison: **persistence and using code as a tool**, exactly what Torvalds wrote about yesterday ("the model gives up before it has exhausted the approaches").

misc: @garrytan on Conductor Cloud - 38k views, "no need to keep the laptop open", the form factor of remote sessions · @garrytan on the loop of working with reality - "Form a

view. Turn it into an artifact. **Put it in contact with reality. Read the result without self-deception.** Revise and run again" · @emollick on consumer agents 35k views - frontier models are good at "**irregular, draining tasks**": a medical bill, a card dispute, tedious forms · @emollick on the shift in discourse - "we've swung too far toward 'AI's impact will show up in years'" · @emollick on lab announcements - asks for "Draw a unicorn in TiKZ" instead of vague posts · @ycombinator: 20+ YC startups are publishing research at NeurIPS/ICLR/ICML, blog · @paulg "How Universities Should Prepare Founders" - 60k views, essay · @blader on gpt-image-2 - "a very tasteful designer": looping a generated design back in beats asking for a design from scratch · @amasad on a Replit milestone 26k views · Your executable is a SQLite database - 508 points, thread · LLMs could take over the host through holes in inference engines - 111 points, thread: the attack targets **vLLM/SGLang themselves**, so "the right harness" is no defence here, a VM is · seL4 proofs completed on AArch64 - 176 points · Sergey Levine on the current state of humanoid robotics - the only fresh Substack in the window

---