DISPATCH 105 · 2026-08-03

GSV THE ADVISORY CITED A FUNCTION NOBODY HAD WRITTEN

The Weights Are Due Next Week

2026-08-03 / 00:22:57

> *"The advisory blamed a function nobody had written yet, and pointed at line numbers in a file that stopped before it got there."*
>
> — from this episode's transcript

- Lenar Kess
- Damra Vol

Two open-weights announcements in the same few hours, and neither one has shipped weights yet. Underneath that, a stranger day: a security registry carrying vulnerability reports for code that was never written, and a research paper whose headline says close to the opposite of its abstract.

- Alibaba announces Qwen3.8-Max — pitched at coding and agentic "cowork", with open weights promised and a 27-billion-parameter sibling due next week. The Hacker News thread hit 694 points, mostly arguing about the smaller model.
- Susan Zhang's hands-on probe — the flagship "has the formatting style of Sol and the reasoning laziness of Opus." One question, not an eval, but from someone who ships models.
- Nathan Lambert and Simon Willison on the cadence — this is the release rhythm now, not a turning point.
- MiniMax opens H3's weights, with a stated hardware floor: verified on a single RTX 5090. The capability claims are the vendor's; nobody outside has published throughput.

- Vaibhav Srivastav moved a daily structured-output job off a frontier model with no quality drop and a roughly twenty-five-fold price cut. David Crawshaw and Jason Calacanis report the same about their own work.
- JFrog: six critical SQLite CVEs describe code that doesn't exist — a function added mid-2025 blamed in version 3.41, line numbers past the end of the file, proof-of-concept payloads that crash nothing, and none of it on SQLite's own advisory page.
- "Agentic Method for Deterministic Validation of Legacy Code Migration" — the paper behind the "bugs included" headline is a test-synthesis method reporting 91.90% branch coverage and exact parity with the COBOL reference. Preserving the bugs is the design goal.
- Gergely Orosz on a GitHub feature that took more than ten months with capable agents available to everyone on it.
- Susan Zhang alleges the explorative-modeling pretraining axis was plagiarized, and in a follow-up says the mode-forcing motivation was lifted too. Her word for the 34-page preprint: "pure claude slop."
- Cornelia Davis of Temporal on why nobody implements MCP Tasks, and the MCP Apps talk on servers returning rendered interfaces instead of text.
- An unconfirmed r/Anthropic report of phantom usage charges tied to stale tokens during a July 29–31 routing incident, and the EU AI Act enforcement date arriving.

SEGMENTS

Transcript

1. Lenar Kess 00:00:04

   Alibaba posted a model announcement a little after seven last night, Pacific time. By the time the West Coast woke up this morning, the Hacker News thread about it was sitting at 694 points and 347 comments, which is a serious thread by any measure. So you'd assume the argument is about the flagship they announced. It mostly isn't. Scroll that thread and the centre of the conversation is a smaller model in the same family that nobody can download yet. [pause] That distance — between the model a company puts in the headline and the model people want on their own hardware — is where today's story sits.

   - x.com
   - qwen.ai
   - x.com
   - x.com
   - x.com
   - x.com
   - x.com
   - x.com
   - x.com
   - x.com
   - x.com
   - x.com
   - reddit.com
   - research.jfrog.com
   - github.com
   - arxiv.org
   - x.com
   - x.com
   - x.com
   - youtube.com
   - youtube.com
   - reddit.com
   - reddit.com

- reddit.com
- euronews.com
- x.com

2. Damra Vol 00:00:40

   The 27-billion-parameter one. Dense, not a mixture of experts, and that's exactly the size where the arithmetic gets fun — that's a model you can quantize down and fit on a card you already own, or on a laptop with enough unified memory. The Max model is an API product. The 27 billion is the one that changes what somebody can do in their kitchen.

3. Lenar Kess 00:01:03

   Right. So here's what was announced. Qwen3.8-Max is Alibaba's most capable model to date, and they've pitched it specifically at coding and at what they're calling cowork. That means agentic use and long-running sessions — the model doing things rather than answering things. Alongside it came two commitments: open weights for Max itself, and Qwen3.8-27B open-weighted next week. <emphasis>Next week.</emphasis> Not today.

4. Damra Vol 00:01:31

   Which is the bit I'd hold onto. Nothing has shipped. A dated promise isn't a download. We've been through enough of these to know that "next week" has a range on it, and that the terms matter as much as the timing — a weights drop under a restrictive licence reads very differently from one under Apache.

5. Lenar Kess 00:01:49

   Fair. Though the promise itself is something — a company marketing a model as frontier-class and saying it'll open the weights is a different posture from a company that never says it. Let me get to the reactions, because the useful ones came from people who actually poked at it. Susan Zhang got hands on it within the hour and reported that it — quote — has the formatting style of Sol and the reasoning laziness of Opus.

6. Damra Vol 00:02:14

   [chuckle] That's a brutal little sentence, and it's more specific than it sounds. She's saying the output surface looks like one lab's model and the effort profile looks like another's. That's the kind of read you only get from someone who's spent years staring at model outputs — she worked on pretraining, she knows what those signatures look like. But it's one probe. She asked it a math question. That's a vibe check, not an evaluation, and I wouldn't want anyone quoting it as a benchmark result tomorrow.

7. Lenar Kess 00:02:43

Agreed, and she didn't present it as one. Nathan Lambert shrugged — his line was that it's another week and another frontier open-weight model. Simon Willison put it next to the announcements from the last few weeks and treated it as part of a sequence rather than a break in one.

8. Damra Vol 00:02:58

   That shrug is about right. Six months ago an announcement like this would've been the whole week. Now it's Monday. The interval between these things has collapsed, and the people whose job it is to be excited about model releases have run out of excitement to spend.

9. Lenar Kess 00:03:13

   Then we've got a video model with open weights and a stated hardware floor. After that, three practitioners reporting from their own work that the open-versus-frontier distinction has stopped mattering to them. There's a security research post about vulnerability reports describing code that was never written, and a paper whose Hacker News headline says close to the opposite of its abstract. Then an attribution fight in pretraining research, and the Model Context Protocol growing two limbs it hasn't had until now.

10. Lenar Kess 00:03:44

    Within a couple of hours of the Qwen post, MiniMax announced open weights for H3, their video model. Simon Willison noticed both in the same window and said so. The detail I'd report from the MiniMax thread is the hardware. They say H3 has been verified running on an RTX 5090 and an RTX 6000.

11. Damra Vol 00:04:06

    A 5090 is a card a person can buy. That's the whole claim, and it's a big one if it holds. A video model that runs on one consumer graphics card means the cost of making moving images stops being a per-second API meter and starts being electricity plus the card you already bought. Anyone doing B-roll, storyboards, pre-visualisation, the middle stretch of a video project nobody puts in the showreel — that changes what they can afford to try.

12. Lenar Kess 00:04:35

    With the obvious caveat that "verified to run on" is the company saying it. There's no independent throughput number anywhere I've seen. Running isn't the same as running usefully — nobody's published seconds of video per minute of wall clock on that card.

13. Damra Vol 00:04:51

And that gap'll close within about seventy-two hours, because the local-model community is good at exactly one thing, which is putting a number on a vendor claim. Somebody will post frames per second on a 5090 with a quantized build before the end of the week. [pause] The other item in their post is omni-reference — holding a character or an object consistent across shots. If that works, it's more interesting than the resolution numbers, because consistency across cuts is what's kept generated video out of anything narrative.

14. Lenar Kess 00:05:23

    Yesterday we mentioned H3 as a release note. Today there's an actual weights announcement with a named piece of silicon attached to it, which is a different item. Tiezhen Wang flagged the listing going up. I'll take the throughput number when somebody measures it.

15. Lenar Kess 00:05:39

    Separately, and over roughly the same day, three people who build things said versions of the same thing about their own work. Jason Calacanis posted that the difference between the open-source models he uses and the frontier models is already negligible for what he does. David Crawshaw — who built Tailscale, and who writes a lot about agents now — argued that because machines write the code, closed tools like Claude Code are less desirable than open source the machine can personalize.

16. Damra Vol 00:06:08

    Crawshaw's argument is the one with real structure in it. He's arguing about reachability rather than quality. If a machine is doing the editing, then the property you want most is that the machine can get inside the tool and reshape it — and a closed product is precisely what it can't get inside. That inverts the usual case for a polished proprietary tool, because polish was for humans.

17. Lenar Kess 00:06:32

    The one with numbers on it is Vaibhav Srivastav's. He moved a daily structured-output task from Sol to Luna, reported no drop in performance, and the price went from five dollars per million input tokens and thirty per million output, down to twenty cents and a dollar twenty.

18. Damra Vol 00:06:49

    So call it twenty-five times cheaper on both sides, on a job he runs every day. That's the kind of number that changes what you're willing to run — not because you save money on the existing job, but because at that price you start running it on everything instead of on the subset you could justify. [pause] Though let me be precise: structured output is about the easiest case there is. Fixed schema, narrow task, verifiable result. It's the workload most likely to survive a downgrade.

19. Lenar Kess 00:07:18

    Take the correction. None of these three are benchmarks. They're three people describing their own workloads, and their workloads aren't the hard ones. There's also a post on the local-model subreddit about KAT Coder 2.5 doing well at modifying functions inside a large undocumented codebase, which is a harder test than structured output and still one person's report.

20. Damra Vol 00:07:42

    And I'd resist stacking these four things into a verdict about open weights winning. What they tell you is narrower and more useful: for a growing set of ordinary production jobs, the frontier premium has stopped buying anything. That's a long way from parity at the top, and the people saying it aren't claiming parity at the top.

21. Lenar Kess 00:08:01

    Here's the item from today I hadn't seen a version of before. JFrog's security research team published a post this morning about six critical vulnerability reports filed against SQLite. Common Vulnerabilities and Exposures entries — CVEs — the formal identifiers that downstream security tooling treats as ground truth. JFrog's finding is that the vulnerabilities don't exist.

22. Damra Vol 00:08:26

    Give me the evidence, because "doesn't exist" is a strong claim about something carrying an official identifier.

23. Lenar Kess 00:08:32

    The evidence is very concrete, and it's what makes the post good. One advisory blames a specific internal function for computing expression operands in SQLite version 3.41. That function wasn't in 3.41 — it went into SQLite in the middle of 2025, years after the version being blamed. Another advisory cites specific line numbers in the JSON source file. In 3.41 that file is 2,706 lines long, and the cited lines are past the end of it. They ran the proof-of-concept payloads. Nothing crashed. And none of these appear on SQLite's own advisory page.

24. Damra Vol 00:09:12

    [tsk] So the advisory blamed a function nobody had written yet, and pointed at line numbers in a file that stopped before it got there. That's not a subtle error. That's what you get when something generates a plausible-shaped security advisory without ever opening the source.

25. Lenar Kess 00:09:28

JFrog say as much. They ran the advisories through GPTZero and reported that all of them read as machine-generated. And the scale is larger than the six — in the same repository of advisories, they say 54 were completely fabricated, and one contained a real bug wrapped in unverified metadata.

26. Damra Vol 00:09:48

    One in fifty-five. And that one's the problem. The fifty-four aren't. If it were all noise you could throw the whole feed away. One real finding buried in fifty-four fabrications means somebody has to read all of them to find it, and that's the exact cost the registry exists to avoid. The Hacker News commenters went straight there — the worry in that thread is that fake entries make the real ones harder to triage, and triage budget is a fixed human quantity.

27. Lenar Kess 00:10:16

    What makes this different from the model-reliability stories we've been running is where the output ended up. This isn't a chatbot being wrong in a chat window. A CVE identifier propagates. It goes into scanners and dependency checkers, and from there into compliance reports and procurement questionnaires. Somebody's automated policy is going to flag SQLite 3.41 as critically vulnerable on the basis of a function that's never existed in that version.

28. Damra Vol 00:10:45

    And the supply is about to increase. There was a Show HN in the same hour — twelve points, barely noticed — for a local pentesting agent that runs on a smartphone. That's a small artifact and I won't inflate it. It runs into the same pipe, though: the cost of producing a security finding is falling fast, and the cost of verifying one hasn't moved at all. The registry was designed around the assumption that filing a report was expensive enough to be a filter.

29. Lenar Kess 00:11:12

    Read JFrog's post directly rather than through the headline, because the per-advisory table is where the case actually gets made. Six identifiers, each with the specific thing wrong with it.

30. Lenar Kess 00:11:24

    Staying with things that aren't what the headline says. There's an arXiv paper on Hacker News today under the title "AI migrated legacy COBOL programs to Java, bugs included." Fifty-two points, thirty-nine comments, and the comments read it as a debunking — agents can't really do migration, they just carry the mess across. I went and read the abstract. That's not what the paper says.

31. Damra Vol 00:11:48

<higher-pitch>Go on.</higher-pitch>

32. Lenar Kess 00:11:49

    The paper's called "Agentic Method for Deterministic Validation of Legacy Code Migration," by Andras Ferenczi, Jordan Docherty, Mariya Bessonov, Matthew Findlay and Krishna Lingamneni. It's a proposed method for validating a migration. They call it the Locksmith Loop. You stand up two runtime environments — the original COBOL and the generated Java — instrument both with mocks, run them off-mainframe on ordinary hardware, and then run an agentic loop that searches for inputs which push execution into unexplored branches.

33. Damra Vol 00:12:24

    So the agent is hunting for the inputs that prove the Java behaves like the COBOL. It's interrogating the migration rather than producing it.

34. Lenar Kess 00:12:33

    Exactly that. And when the search stalls, an analyzer identifies what they call a Locked Paragraph — a condition that's blocking deeper exploration. Three case studies, programs ranging from 430 to just over 4,000 source lines. Near-complete coverage on the two open-source programs, and 91.90% branch coverage on the internal production-like one. Then there's the result they lead with: the generated Java matched the COBOL reference under deterministic parity checks in all accepted test cases.

35. Damra Vol 00:13:06

    Then "bugs included" describes the specification, not a failure. If the mainframe program has been miscalculating something since 1987 and forty downstream systems have been compensating for it all along, a migration that fixes the bug is a migration that breaks production. Parity means parity. You want the bugs. You port them deliberately, and then you fix them afterwards, on purpose, with the downstream consumers identified first.

36. Lenar Kess 00:13:34

    Which is the part everybody underestimates when they price one of these projects.

37. Damra Vol 00:13:39

    Behaviour-preserving is the hard requirement in migration work, and a method that can demonstrate it on commodity hardware without a mainframe in the loop is a real contribution. Whether 91.90% branch coverage is enough to sign off a production cutover is a separate and much less comfortable question.

38. Lenar Kess 00:13:58

    The commenters did raise one objection the paper doesn't answer, and it's the right one. A COBOL program isn't a standalone artifact. It sits inside job control language, CICS transaction handling, mainframe sort processors — and none of those have a Java equivalent to migrate to. You can prove the program is faithful and still have nothing that runs.

39. Damra Vol 00:14:21

    Which is why these projects take years, and why the demo is always the easy part. Gergely Orosz posted something adjacent yesterday — a GitHub feature that's taken more than ten months to ship, at a company where everyone working on it has capable agents available. He isn't claiming the agents failed. He's pointing at the fact that calendar time on a complicated project in a mature codebase hasn't compressed, whatever happened to typing speed.

40. Lenar Kess 00:14:48

    Two separate observations, and I'll keep them separate. But if you're wondering why the enterprise migration pitch keeps outrunning the enterprise migration results, the substrate question is most of the answer.

41. Lenar Kess 00:14:59

    Susan Zhang again, earlier this morning, on something entirely different. She posted that the work presenting explorative modeling as a third axis of pretraining looks to her like plagiarism. In a follow-up she said the motivation for mode forcing was lifted as well, and she pointed at a specific 34-page preprint, which she described as — her phrase — pure claude slop.

42. Damra Vol 00:15:23

    Every sentence of that has to stay attached to her name. This is one researcher on X, there's no response from the other side anywhere in today's material, and neither of us can adjudicate who had which idea first. What I'll say is that she's not a random account — she has pretraining credentials, she worked on the OPT models at Meta, and she named specifics rather than gesturing.

43. Lenar Kess 00:15:48

    Which is why it's reportable. An unattributed accusation is gossip. A named researcher pointing at a named document with a named mechanism is a claim somebody can answer.

44. Damra Vol 00:15:58

    But the detail I keep circling is the insult. "Pure claude slop," about a 34-page preprint. Ten years ago the way you dismissed a paper you thought was thin was to attack the method or the evidence.

Now there's a shorter move available — you say a model wrote it — and it functions as a complete argument without anyone having to demonstrate anything.

45. Lenar Kess 00:16:20

    And it's unfalsifiable in both directions. The accused can't prove a human wrote it. The accuser can't prove one didn't.

46. Damra Vol 00:16:28

    Which makes it a very effective weapon and a very poor argument. I don't think Zhang is wrong to be angry — if someone took your idea and dressed it in thirty-four pages of generated prose, anger is the correct response. But the accusation has got cheap enough that it'll get used by people with much worse cases, and every field that adopts it loses the ability to tell the two apart.

47. Lenar Kess 00:16:51

    Two talks from the AI Engineer conference went up within hours of each other yesterday, and together they describe the Model Context Protocol — MCP, the standard for how agents talk to tools — picking up two things it's lacked. Start with the less glamorous one. Cornelia Davis, a technologist at Temporal, gave a talk explaining why essentially no agent implements MCP Tasks.

48. Damra Vol 00:17:16

    Tasks covers tool calls that don't return immediately. You invoke something, you get a handle back, and the work continues — for minutes, for hours, across a human approval step in the middle.

49. Lenar Kess 00:17:27

    Right. And Davis's account of why nobody has built it is specific rather than hand-wavy. The version from November is stateful. It has a task-list endpoint with no filtering, so a server holding millions of concurrent tasks has to hand back all of them. And getting human input into a running task requires holding open a persistent connection, so every reconnection turns into a state-recovery problem.

50. Damra Vol 00:17:52

    That's a spec written by people imagining tens of tasks. Any durable-work system that's survived contact with production ends up in the same place — you store the state, you poll or you subscribe, and you never assume the connection outlives the work. Temporal exists because that problem is hard. Of course the person who turned up to explain the gap works there.

51. Lenar Kess 00:18:15

She demonstrated it with a purchase-order workflow, backed by Temporal for the state and the parallel orchestration, walking through the lifecycle states — working, input required, completed — and the signal that lets a human approval reach a long-running backend process. The July revision goes stateless. Tasks become an optional extension of a modular core, the task-list endpoint goes away, and the input tunnel gets replaced by explicit client-to-server update calls.

52. Damra Vol 00:18:44

    Which is the correct redesign, and her closing caution holds up: making the protocol stateless doesn't make the problem go away, it relocates it. Somebody still has to build the durable orchestration underneath. The spec just stops pretending it can do that for you.

53. Lenar Kess 00:19:00

    The other talk is MCP Apps, from Ido Salomon and Liad Yosef, who maintain the specification. The idea is that a server can return an interface rather than text. It sends HTML or components through ordinary MCP resources, and the host renders them in a sandbox inside Claude, VS Code, Cursor or ChatGPT. Their statement of the problem is that today a service handing its data to a model gets flattened into a text database. No layout, no brand, no interaction.

54. Damra Vol 00:19:32

    And the mechanism is where it gets strange. The rendered widget is isolated, and when you click something inside it — a row in a funnel dashboard, a track in a list — that click serializes into an event that goes back through a callback to the model. The model stays in charge of the flow and decides what to call next. So the interface acts as an input device for the agent rather than an escape hatch out of the conversation.

55. Lenar Kess 00:19:58

    With three defined tiers of how much control a host hands over: notify only, prompt the host to run something, or hand the turn fully to the chat. It's being developed in the open under the MCP steering committee, with Anthropic and OpenAI both contributing directly.

56. Damra Vol 00:20:16

    The speakers also project it becomes the global standard for distributing interfaces by 2026, which is their projection rather than a fact — they maintain the spec, they're supposed to believe that. What I'd say more modestly is that if this works, the unit of software distribution stops being a page or an app and starts being a widget that shows up inside somebody else's assistant. That's a strange thing to build toward, and it's still experimental.

57. Lenar Kess 00:20:43

Two other items before we stop. There's a post on the Anthropic subreddit describing a billing bug — the account says that during a routing incident from July 29th through the 31st, revoked authentication tokens kept being billed against Pro and Max plans, and reports a card charged more than four hundred and eighty dollars, no response from support, and being banned from the Discord for raising it.

58. Damra Vol 00:21:07

    One account, unverified, and there's no statement from Anthropic anywhere in today's material. I wouldn't repeat the dollar figure as established. What makes it more checkable than the usual complaint is the mechanism and the date window — stale tokens, a specific three-day incident. That's the kind of claim that either matches other people's invoices or it doesn't, and we'll know fairly quickly.

59. Lenar Kess 00:21:30

    There's a separate thread arguing the current model series is unreliable in a specific way, and a lighter one about deleting your project instructions file — apparently on the suggestion of an Anthropic engineer — with at least one person reporting they tried it and preferred the result.

60. Damra Vol 00:21:46

    [chuckle] I have some sympathy for that. A lot of those files have accreted into a wall of instruction that the model half-follows and the human never rereads.

61. Lenar Kess 00:21:56

    The last item is a date. The EU AI Act obligations we went through in detail yesterday became enforceable. Armin Ronacher's entire response was one line — I'm really curious if people will comply with this. The next concrete thing in that story is a first enforcement action against a named model provider, and until that happens nobody has to answer his question.

62. Damra Vol 00:22:19

    Meanwhile the two model announcements that opened the show both promised weights they haven't delivered, and the security registry is carrying six vulnerability reports for code that doesn't exist. Those are unrelated stories and I'm not going to pretend otherwise. But they rhyme in one narrow way, which is that the claim arrived well ahead of the artifact.

63. Lenar Kess 00:22:39

    The artifact I actually want is the Qwen 27-billion-parameter weights, and the licence they come under. If that shows up next week the way it was promised, the local-model side of this gets a better

week than the flagship announcement gave it.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic