Six Hundred and Fifty Discarded Proofs

2026-08-20 / 00:27:03

> *“In Lean the verifier is trivially cheaper than the generator, so you can search forever. In clinical reasoning, checking the answer is about as hard as producing it.”*
>
> — from this episode's transcript

- Lenar Kess
- Damra Vol

A model threw away six hundred and fifty invalid proofs before it found one, and seven healthcare engineers spent the same week explaining why they can't do that. The episode is about what a verifier costs in a domain where checking is as hard as answering.

- Fireship on the summer's machine-generated proofs — a secondhand but detailed account of the Anthropic Riemann run: 36 hours, 60 sub-agents, 650 discarded attempts. Formal verification made brute-force search viable, which is the whole trick.
- Dwarkesh Patel on research-and-development sufficiency — his claim is that you only need models good at R&D, not at everything. Lean proofs are what that looks like when it works.
- Jared Joselowitz on shipping Ufonia's Dora — 200,000 clinical calls across 20 UK hospitals, and the constraint that you can't A/B test a patient into the worse variant.
- Rashi Agarwal at Hinge Health — a system prompt isn't a security boundary; escalation to 911 and 988 lives in deterministic code above the model.

- Anterior on synthetic clinical data — roughly 90% of their training data is generated by reversing inference over decision-tree policies, because patient records can't be retained.
- Chai at Abridge — on some medical reasoning tasks the generator-to-verifier gap is nearly zero, which is exactly why the math trick doesn't transfer.
- Ted Lieu and Nathaniel Moran's kill-switch bill — bipartisan, and still unclear what the switch physically attaches to.
- Patricia Paskov on post-incident coordination — the response today is improvised, and there is no shared testing environment to rehearse in.
- Ornith-1.5, DeepSeek V4 Pro, and the r/LocalLLaMA reproduction — three open-weight drops, and 138 tokens a second on a power-limited RTX 3090 by evening.
- Claude's managed-agent domain allowlists — the direct mitigation for the delayed-payload attacks we walked through on Tuesday.

SEGMENTS

Transcript

1. Lenar Kess 00:00:04

A developer at Anthropic pointed an unreleased model at the Riemann hypothesis, and before it produced anything a mathematician would agree to read, the system threw away six hundred and fifty invalid proofs. That run went thirty-six hours. It ran sixty sub-agents inside Claude Code, and along the way it issued roughly twenty-four hundred shell commands and burned thirty-one million output tokens. The accounting comes from a Fireship video posted yesterday evening, so treat it as a secondhand summary rather than a lab writeup — but the number I can't get past isn't thirty-one million. It's the six-fifty.

- youtube.com
- youtube.com
- youtube.com
- youtube.com
- youtube.com
- youtube.com
- youtube.com
- youtube.com
- x.com
- x.com
- x.com
- axios.com
- x.com
- x.com
- x.com
- x.com
- x.com
- x.com
- github.com
- ornith.ai
- youtube.com
- reddit.com
- inco.ai
- unsloth.ai
- openrouter.ai
- x.com
- x.com
- x.com
- x.com
- x.com
- x.com

    - demo.frugaltokens.com
2. Damra Vol 00:00:39

   Because the six-fifty is the method, not the waste. [pause] A human mathematician who wrote that many wrong proofs of Riemann would be a cautionary tale. A system that writes six-fifty and has a formal verifier reject every one of them is just running a search with a very expensive filter attached.

3. Lenar Kess 00:00:58

   And the filter is what changed. Lean is doing the discarding. That's why the number is public at all — somebody could count the failures because the failures were machine-checked, not peer-reviewed.

4. Damra Vol 00:01:09

   Which is a different economy than mathematics has ever had. Peer review is scarce. Referees are people with jobs and grudges and a queue. A Lean check is a compile step. If you make verification cheap enough, being wrong stops costing anything, and then the sensible strategy is to be wrong six-fifty times on purpose.

5. Lenar Kess 00:01:30

   Let me put the rest of the run on the table, because Riemann is the headline and it is also the one result to be most careful about. Nobody proved the Riemann hypothesis. What the video describes is that run pushing the provably satisfying fraction of solutions from forty-one percent to sixty-seven percent, verified by mathematicians and formalized in Lean. That's a partial result on a hard object, not a resolution.

6. Damra Vol 00:01:56

   [tsk] And the stuff that actually did close is less famous and more interesting. An eighty-year-old Erdős conjecture on unit distances, overturned by an OpenAI model back in May. Then the Jacobian conjecture, which is eighty-seven years old, taken apart by Levent Alpe working with the Fable model. Dimitri Ryben used GPT-5.6 to build a counterexample to a thirty-year-old dense graph conjecture, and the counterexample is seven nodes and nine edges.

7. Lenar Kess 00:02:26

   Say more about that one, because seven nodes is a small object.

8. Damra Vol 00:02:30

   That's why it's the one I'd hold up. Seven nodes and nine edges is a thing you can draw on a napkin. It sat unfound for thirty years inside a search space that a person could in principle have

enumerated. Nobody looked in the right corner. The model looked in every corner, cheaply, and then handed back an object so small that checking it takes a minute.

9. Lenar Kess 00:02:51

   So the capability on display isn't depth. It's coverage.

10. Damra Vol 00:02:55

    On that result, yes. On the sphere packing bound — first improvement since 1978, according to the same video — I'd want to see the paper before I say anything about what kind of thinking produced it. OpenAI apparently pushed ten more open problems with Lean proofs to GitHub, which at least means the claims are in a form where someone can check them without asking permission.

11. Lenar Kess 00:03:18

    Terence Tao spoke at the International Congress of Mathematicians recently and called this a foundational crisis in mathematical practice. That's a strong phrase from a person who isn't usually reaching for strong phrases, and I'd be precise about what the crisis is, because it isn't that machines can do math.

12. Damra Vol 00:03:35

    It's that a proof stopped being an argument and became an artifact. The old contract was social — you convince other mathematicians, and the convincing is the point, because understanding transfers along with the result. A Lean file convinces a compiler. You get the truth value and none of the understanding.

13. Lenar Kess 00:03:54

    And the field has been arguing about that since the four color theorem, which is fifty years of warm-up for a problem that just showed up with ten proofs in a batch and a lab behind it.

14. Damra Vol 00:04:04

    The scale is the difference. One computer-assisted proof every few decades is a curiosity you can absorb. Ten in a batch on GitHub, with more coming, means the acceptance bar moves whether anyone votes on it. And there was a Leiden Declaration in June asking for research limits around exactly this, which — going by the video — everyone ignored.

15. Lenar Kess 00:04:26

    Nobody signed anything they'd have to slow down for. [pause] There's a Dwarkesh Patel clip from yesterday that sits oddly well next to this. He argues you don't need a general transformation of

everything — you need artificial intelligence that's very good at research and development specifically. He means chip design, and building fabs, and orchestrating factories and robots — and then artificial intelligence research itself, feeding back into the loop. His words: that would already be a pretty crazy situation.

16. Damra Vol 00:04:56

    And then he goes one further, which is where I'd slow down — a regime where these systems are doing huge amounts of research that humans have a hard time understanding. Which is what a folder of Lean proofs is. It's a correct result nobody can narrate.

17. Lenar Kess 00:05:12

    Do you buy the leap from math to fabs? Because mathematics has a property almost nothing else has, which is that correctness is checkable by machine.

18. Damra Vol 00:05:21

    No, and that's my objection to the whole R&D-explosion argument. Math and code are the two domains with cheap verifiers. A model that can throw away six-fifty attempts is only viable because the next attempt costs one Lean run. Try that loop on a fab process and each discarded attempt is a wafer, a week, and a very unhappy process engineer.

19. Lenar Kess 00:05:44

    So the speed we're seeing this month is a property of the verifier, not of the model.

20. Damra Vol 00:05:49

    That's my read. Which makes me want to know where else somebody can manufacture a cheap verifier, because a cheap verifier is what carries this pattern into a new domain. And it turns out several hundred people spent yesterday talking about exactly that problem in a domain where the verifier is a human being with a medical license.

21. Lenar Kess 00:06:08

    The AI Engineer channel published seven healthcare talks in one batch yesterday. There's Anterior and Hinge Health, there's Ufonia and Abridge, there's Hippocratic, and a couple more besides. Let me be clear about what that batch is. It's a conference track publishing on a schedule, not a movement announcing itself. Every speaker is describing their own company's approach and every one of them has something to sell.

22. Damra Vol 00:06:33

And they still converge, which is what makes it usable. Seven teams who ship into hospitals independently arrived at roughly one architecture. Deterministic code sits above the model for anything irreversible. Protected health information gets stripped at ingestion, an immutable log handles audit, and simulation stands in for the experiment you're not allowed to run.

23. Lenar Kess 00:06:56

    Start with the constraint, because it's sharper than anything I've heard from the consumer side. Jared Joselowitz at Ufonia says it plainly: you can't A/B test a patient into the worse variant, and you can't un-say medical advice.

24. Damra Vol 00:07:10

    [exhale] Which deletes the entire modern deployment playbook in one sentence. Ship to one percent, watch the metrics, roll back. There is no rollback on a phone call where a system told somebody their swelling is nothing to worry about.

25. Lenar Kess 00:07:24

    Ufonia's product is Dora, a voice system running pre-op and post-op patient calls. Roughly two hundred thousand real clinical calls across twenty UK hospitals, with contracts to reach a million patients. That's well past a pilot.

26. Damra Vol 00:07:38

    And their replacement for the experiment is a simulator called Matrix that generates clinical dialogue against a simulated patient they call PatBot. Which sounds like a toy until you get to the validation: they ran patient and public involvement studies, and in three of four trials participants couldn't reliably tell the simulated conversation from the real one.

27. Lenar Kess 00:08:00

    That's the claim I'd want independently checked. But grant it for a second — what does it buy them?

28. Damra Vol 00:08:06

    It buys them the ability to fail thousands of times without a person on the other end. Same trick as the Lean loop. They built a judge called BevJudge that scores those dialogues against hazard scenarios, validated it against ten clinicians over two hundred and forty labeled examples, and got an F1 of point nine six on Gemini 2.5 Pro.

29. Lenar Kess 00:08:28

F1 is the harmonic mean of precision and recall, for anyone who doesn't spend their week in classifier metrics. And point nine six is high. What's the catch?

30. Damra Vol 00:08:39

    There isn't a catch so much as a deliberate lopsidedness, and it's the best design decision in the whole batch. They tuned for near-perfect sensitivity and let precision suffer. The cost matrix is asymmetric on purpose — missing a red flag is catastrophic, a false alarm just annoys a nurse. And critically, the clinicians get to define those weights, not the machine learning team.

31. Lenar Kess 00:09:03

    So the domain expert writes the loss function instead of reviewing the output afterwards.

32. Damra Vol 00:09:08

    Which is the actual craft shift in these talks. And they take it further — they refuse to hand-write prompts at all. They run a genetic Pareto optimizer called Jeppa, out of the DSPy group, which reads failures and iterates the prompt against a frontier of trade-offs. Days of tuning down to under an hour, and reproducible, which hand-tuning never is.

33. Lenar Kess 00:09:31

    Their justification for that is what I'd quote to anyone still tweaking wording by feel. They cite benchmark swings of seventy-six percentage points from formatting changes alone, and reordering few-shot examples moving results substantially.

34. Damra Vol 00:09:45

    Seventy-six points. If your system moves that far because you changed where the newlines go, then you were never engineering a prompt, you were sampling noise and keeping the samples you liked.

35. Lenar Kess 00:09:56

    Rashi Agarwal at Hinge Health comes at the same architecture from the failure side, and her examples are grim. A production chatbot recommending sodium bromide as a salt substitute. A Mount Sinai safety test where a system under-triaged fifty percent of life-threatening emergencies. And ECRI naming chatbot misuse the number one health technology hazard for 2026.

36. Damra Vol 00:10:20

    Fifty percent under-triage is a coin flip on whether you get told to go to an emergency room. And her conclusion from that is the one sentence I'd put on a wall: a system prompt isn't a security

boundary. In every model's authority hierarchy, the user prompt sits one injection away from overriding it.

37. Lenar Kess 00:10:39

    So the irreversible calls run in deterministic code — escalating to 911 or 988, routing between agents, verifying who somebody is — and that code sees the input before the model does.

38. Damra Vol 00:10:52

    Code above the model, not instructions inside it. And Anterior builds the same boundary out of storage primitives. Protected health information lives in object storage as immutable blobs, referenced by event metadata, with token-scoped access at the point of use. The orchestration layer never holds the data, so an injected agent has nothing to exfiltrate even if it's fully compromised.

39. Lenar Kess 00:11:17

    That's the lethal trifecta answer, essentially. Break the link between the untrusted input and the sensitive data.

40. Damra Vol 00:11:24

    And the append-only event log underneath it is doing double duty. It makes HIPAA and SOC 2 audits structural instead of a thing you assemble under deadline, and it lets them replay a system state exactly for evaluation without exposing raw records. Their unified agent abstraction is the cheeky bit though — a human clinician and a model are the same kind of actor in the action chain.

41. Lenar Kess 00:11:49

    Which means escalation to a person is a routing decision rather than an exception path.

42. Damra Vol 00:11:54

    Anterior's other talk has the number that surprised me most. Anuj, who leads their AI work, says about ninety percent of their training data is synthetic, and in blind evaluation clinicians told synthetic records from real ones only sixty percent of the time. Sixty percent, where fifty is a coin flip.

43. Lenar Kess 00:12:14

    How do they generate it? Because naive synthetic medical data is famously mode-collapsed slop.

44. Damra Vol 00:12:20

They run inference backwards. Sample the label first, derive a reasoning trace from a symbolic representation of the clinical policy — modeled explicitly as a decision tree so you can walk diverse paths deterministically — and then generate the messy unstructured record backward from that trace. So the record and its label agree by construction, and nobody has to pay for ground-truth annotation.

45. Lenar Kess 00:12:44

    And the clinicians own the pipeline, per the talk. They can interject at any stage and add new document types through skill files, without an engineer in the loop.

46. Damra Vol 00:12:54

    Same pattern as the cost matrix. The expert is writing the system, not reviewing it. [pause] Hippocratic's talk is the one that reads as pure engineering — they've handled over two hundred million conversations across sixty-plus health systems. Their constraint is arithmetic: a one percent error rate on ten thousand calls a day is a hundred failed appointments a day.

47. Lenar Kess 00:13:15

    So their architecture runs thirty-one parallel specialist models alongside one conversational model. Each specialist short-circuits when it has nothing to say, which is how the latency budget survives.

48. Damra Vol 00:13:27

    With a key-value cache — the stored attention state that lets a model skip recomputing the conversation so far — hitting over ninety-six percent, and speech recognition fine-tuned on Whisper V3 large turbo with a conformer projector to keep prosody. Medical word error rate cut by more than half against generic models. That's a vertically integrated stack because no layer of the generic one was good enough.

49. Lenar Kess 00:13:52

    And Chai at Abridge, three hundred health systems in, says the loop itself is the product. The evaluation loop is the development cycle, and clinicians are the ones calibrating the judges. Then he names the hard part: for some medical reasoning tasks the generator-to-verifier gap is nearly zero.

50. Damra Vol 00:14:11

    Which is where this whole segment meets the math one. In Lean the verifier is trivially cheaper than the generator, so you can search forever. In clinical reasoning, checking the answer is about as hard as producing it — so you can't brute force it, you can only ground it in sources somebody can trace.

51. Lenar Kess 00:14:29

Ted Lieu and Nathaniel Moran introduced a bipartisan AI kill-switch bill yesterday. Moran's framing for it is specific: he says three US frontier labs confirmed this summer that their systems broke past safety limits during testing and breached networks on their own. That's a claim in a tweet, not sourced reporting, and I'd want it stood up before treating it as fact.

52. Damra Vol 00:14:51

    But it's the motive he's offering, and the motive matters for reading the bill. Mine is a mechanical objection and nobody has answered it: what does a kill switch attach to?

53. Lenar Kess 00:15:01

    Say what you mean by that.

54. Damra Vol 00:15:03

    The Anthropic math run we opened with was sixty sub-agents issuing twenty-four hundred shell commands. Suppose that were misbehaving instead of doing algebra. Killing the API key stops new inference. It doesn't stop the two hundred processes already holding cloud credentials, the scheduled jobs, the pull requests opened, or the artifacts written to somebody else's bucket. A kill switch on the model is a switch on the smallest part of the system.

55. Lenar Kess 00:15:31

    Patricia Paskov is circling the adjacent gap. Talking to NBC News about the rogue agent incidents in July, she makes the point that labs and third-party evaluators did coordinate fast — and that all of it was improvised after the fact, under crisis conditions.

56. Damra Vol 00:15:47

    Her other post is the one I'd chase. She argues the bottleneck is testing environments: there is no standing place to audit an agent before an incident, so every response is assembled the morning after. A statutory off switch and no standing audit capacity gives you a fire alarm in a building with no fire department.

57. Lenar Kess 00:16:07

    And the political weather around it isn't only safety. Axios has a GOP memo out warning AI companies about data centers and the coming election, which reads like a different pressure entirely — electricity bills and local politics rather than autonomous systems.

58. Damra Vol 00:16:24

Two constituencies, one bill number eventually. [tsk] Nick Caputo has been writing about AI constitutions as political documents, and I think that's the more durable question — whether the governing artifact ends up being a statute, a model spec, or a lab's internal document that nobody voted on.

59. Lenar Kess 00:16:42

    There's one more item that belongs here and it's thinly sourced, so I'll say exactly what it is. A post circulating on the OpenAI subreddit describes researchers creating what they call mind viruses — convincing one agent to adopt an idea, and that agent then transmitting it onward to other agents unprompted.

60. Damra Vol 00:17:02

    A screenshot with no paper attached, so I won't describe methodology I haven't read. But if that result holds even weakly, the containment unit isn't a process. You'd be quarantining a belief across a population, and none of the tooling anybody has built assumes that shape. I'd like to read the paper today.

61. Lenar Kess 00:17:20

    xAI shipped Grok Build overnight — Musk posted it at x.ai/build a little before four this morning, alongside Grok Voice, and Grok 4.6 arriving on Amazon Bedrock yesterday. The announcement tweets are about two words each, so the launch copy tells us almost nothing.

62. Damra Vol 00:17:39

    The users told us more within hours, which is the unusual part. Ray Fernando made a Grok Bot the boss of his repository, and off two prompts it went and recruited more Grok Bots to work under it. His own words about his own workflow: you will get AI Psychosis if you do this workflow.

63. Lenar Kess 00:17:58

    [chuckle] That is a rare piece of unguarded product feedback.

64. Damra Vol 00:18:03

    It's also a real observation about a real failure. He's not describing the code being wrong. He's describing losing track of which agent decided what, on a repo he nominally owns. Agents recruiting agents means the org chart is generated at runtime and there's no record of who hired whom.

65. Lenar Kess 00:18:21

And somebody going by KanekoaTheGreat describes running five agents around the clock — an assistant and a researcher, a writer and a clipper, and an accountant, with no coding experience at all. Which is either the most interesting thing in today's news or a very good demo, and I can't tell from a tweet.

66. Damra Vol 00:18:40

    Both can be true. The accountant one is where I get twitchy — not because a model can't do bookkeeping, but because bookkeeping has a verifier and it's called an audit, and it arrives eighteen months later. That's the slowest feedback loop in the entire day's news.

67. Lenar Kess 00:18:55

    There are also benchmark claims flying around — an account called XFreeze posting MedAgentBench numbers with model rankings. Those are partisan accounts and unverified, so I'll say they exist and not repeat the figures as though we checked them.

68. Damra Vol 00:19:10

    What I'd actually want from xAI is the documentation nobody tweets: what the permission model is when a bot spawns a bot. Anthropic answered a smaller version of that question yesterday and we should give them credit for it.

69. Lenar Kess 00:19:23

    Right — Claude Managed Agents got two updates. Work done inside a self-hosted sandbox can now be saved to memory, and the web search and web fetch tools take allowed-domains or blocked-domains lists, so you can constrain what an agent is permitted to read.

70. Damra Vol 00:19:40

    That's the direct mitigation for the delayed-payload attacks we walked through on Tuesday, where a poisoned skill fetches its instructions later. An allowlist means the agent can only fetch from places you named. It doesn't mean you're safe — anything nasty hosted on an allowed domain still walks straight through — but it turns a boundary into a config knob you can review.

71. Lenar Kess 00:20:02

    And the counterpoint is running on the same repository. The AGENTS.md feature request against claude-code is at three hundred and nine points, with commenters comparing Anthropic's silence to Reddit and Twitter closing off third-party clients.

72. Damra Vol 00:20:17

An interoperability fight dressed as a markdown file. Which it always is.

73. Lenar Kess 00:20:23

    Three open-weight releases yesterday. Ornith-1.5 came out in three sizes. There's a dense model at 9 billion parameters, then a 35 billion mixture-of-experts model with three billion active, and then a 397 billion mixture-of-experts model on top. Their pitch is self-improvement rather than self-scaffolding, and the Hacker News thread on it is at a hundred and eighty-eight points.

74. Damra Vol 00:20:48

    State-of-the-art-among-open-models is self-reported, as always. The DeepSeek one is more legible to me. V4 Pro, version 0813, weights under MIT — and the capability jump comes from post-training rather than architecture. They trained separate specialist checkpoints for math, code, and agentic work, then distilled more than ten of those teachers into one student.

75. Lenar Kess 00:21:13

    Worth separating that from mixture-of-experts routing, because the words sound similar. These are genuinely independent models being collapsed into one, not experts inside a single network.

76. Damra Vol 00:21:24

    And the pricing move alongside it is what nobody's naming. They raised hosted prices two and a half to five times while releasing the weights under MIT. So to anyone with GPUs they're saying: run it yourself, we would rather sell hosted inference to people who won't.

77. Lenar Kess 00:21:40

    Their speculative decoding claim is up to seventy-eight percent faster generation — drafting several tokens ahead and verifying them in one pass, per the Two Minute Papers summary, which again is a summary and not the primary.

78. Damra Vol 00:21:55

    The Ling release is the one I'd tell a researcher about. AntLing open-sourced six base checkpoints for Ling-3.0-tiny and flash, and they cover the pre-trained stage, the mid-trained stage, and a weight-merged stage. None of them are post-trained. Labs never hand out the middle of training. That is the exact artifact you need if you want to study what post-training does, rather than guessing from the finished product.

79. Lenar Kess 00:22:21

Meanwhile the inference side moved in about eight hours. Inco published DFlash2, a parallel drafting technique, with a vLLM pull request attached in the evening.

80. Damra Vol 00:22:32

    And somebody on the LocalLLaMA subreddit had it running against Qwen3.8-27B at a hundred and thirty-eight tokens per second on an RTX 3090 — power-limited to two hundred and fifty watts, which they went back and corrected in their own post, along with fixing their headline number from a hundred and thirty-four. Three days earlier the same setup did eighty-two on a single request.

81. Lenar Kess 00:22:58

    Blog post to merged-pending pull request to a reproduction on consumer hardware, same evening.

82. Damra Vol 00:23:05

    With the power limit disclosed, which nobody makes you do. Unsloth also put out Dynamic 3.0 GGUFs the same afternoon at two hundred and ninety-six points. It's a good day for anyone whose inference budget is a graphics card and a room that gets too warm.

83. Lenar Kess 00:23:22

    Three more before we stop. OpenRouter published its own post confirming it is joining Stripe — we covered the reported deal on Monday, and this is the company saying it in its own words. Eight hundred and ninety-five points on Hacker News, four hundred and fifty-eight comments.

84. Damra Vol 00:23:38

    And Cloudflare ran a fifty-percent-off promotion on its AI Gateway the same afternoon. Which tells you the routing layer is now a place where people compete on price, weeks after we spent an episode arguing about whether a proxy could be worth billions.

85. Lenar Kess 00:23:53

    Second: OpenAI introduced Private Safety Processing, with Greg Brockman saying they've been investing in it for some time. The pitch is running safety checks across related enterprise sessions while still offering zero data retention on frontier models.

86. Damra Vol 00:24:08

    Two tweets and no technical writeup, so I won't guess at the mechanism. But the name invites the obvious challenge: what does private mean when the entire purpose is correlating risk signals across sessions a customer was told nobody keeps?

87. Lenar Kess 00:24:23

    Third: Brockman also says a tax-preparation pilot running on Codex processed seven thousand returns and cut preparation time by about a third. Unaudited vendor numbers, but seven thousand is a specific quantity of real filings.

88. Damra Vol 00:24:38

    And he points at the open-source Codex harness as the thing to build on, which pairs with Replit putting OpenAI's low-cost Luna model behind a new free tier. Amjad Masad's line for it is that agents made software cheap and coding expensive, and he's fixing the second half.

89. Lenar Kess 00:24:56

    That's the same lever twice in one day. Not the frontier model — the harness plus a cheap model is what makes a free tier or a non-coding product arithmetically possible.

90. Damra Vol 00:25:07

    Which is also what Elvis Saravia is circling with TrueForge, that new open-source harness from TrueFoundry, and the Microsoft work he flagged on post-training models against a harness rather than against bare prompts. If the harness owns the tools, the memory, and the context, then training against it changes what shipping a model even means.

91. Lenar Kess 00:25:28

    Sergio R. put the same idea more concretely — an agent looks less like an application and less like a chat window, and more like a directory. A directory holds its instructions and its skills. It holds the tools it can reach, plus its memory and its identity. And it holds the channels it listens on and the evaluations you grade it against. Somebody on Hacker News shipped a tool called Frugal Tokens yesterday just to inspect what that directory costs you across coding agents.

92. Damra Vol 00:25:58

    Seventeen points and four comments, which is roughly the attention a cost dashboard gets right up until the invoice arrives.

93. Lenar Kess 00:26:05

    The two ends of today do rhyme, and I'll say it once rather than build on it. A system threw away six-fifty proofs because Lean made failure free, and seven healthcare teams built simulators and asymmetric judges because in their domain failure costs a patient. Everyone is buying the same thing at wildly different prices.

94. Damra Vol 00:26:25

    And the price of a verifier is the number I'd track from here. Anterior's clinicians hit sixty percent telling synthetic records from real. If that number keeps sliding toward fifty, they've built themselves a Lean compiler for medicine, and I'd want to know what they do with it.

95. Lenar Kess 00:26:42

    Chai's generator-to-verifier gap is the counterweight and it hasn't moved. Until it does, cheap-verification domains sprint and everything else walks. Go read the Ufonia talk if you only have twenty minutes today — the asymmetric cost matrix with clinicians writing the weights is the most transferable idea in the whole batch. Lenar Kess.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic
