A Lab Stops Its Own Biggest Run

2026-08-19 / 00:21:28

> *“Twenty percent of research inference compute spent watching your own models think is not a press release. That is a budget line, and budget lines are the part of a safety commitment you can actually check.”*
>
> — from this episode's transcript

- Lenar Kess
- Damra Vol

OpenAI says it has paused reinforcement learning training on its latest deployment-bound models while new security and monitoring requirements catch up. On the same day, an outside group graded every frontier lab on whether its control practices are actually implemented, Anthropic disclosed three unreleased internal models, and a governor signed data center siting rules while Nvidia backed a 4.25 gigawatt build next door.

- OpenAI announced a temporary pause on reinforcement learning runs for models intended for deployment — the pause is specific to that phase, not to all training.
- Pacing model development in an era of cyber-critical capabilities is the policy post underneath the announcement, and it names cyber capability as the pacing constraint.
- Greg Brockman and Sam Altman gave two different accounts of the same decision — confidence setting the pace versus capability outstripping safety.

- Max Zeff reports the next frontier model, Astra, is held up on security and alignment work, and the largest planned run remains on hold.
- Ethan Mollick flagged the disclosed figure of roughly 20% of research inference compute going to chain-of-thought monitoring.
- Steven Adler's Guidelight scorecard grades frontier labs on control practices and finds all of them at most partially implemented, on public evidence alone.
- Anthropic's protein design campaign reports a 35% binder success rate, and Ravid Shwartz Ziv accounts for the 30k-token expert prompt and roughly 12,500 H100-hours behind it.
- Governor Josh Shapiro signed data center standards for Pennsylvania the same day a 4.25 gigawatt Ohio build was reported at 105 billion dollars.
- Cerebras CS-4 and Etched's 700 million dollar raise arrive as average token prices halve.
- LangSmith Tuned Evaluators and fx, a 6.3 mebibyte coding agent in Zig, both aim at agent work from opposite ends of the size range.

SEGMENTS

Transcript

1. Lenar Kess 00:00:04

   Picture the position for a second. You have one of the largest training clusters on the planet, a competitor shipping every few weeks, and a model your own executives keep hinting at in public. What would it take for you to stop? Not stop everything — stop one specific category of run, on purpose, and then publish the fact. Yesterday afternoon, OpenAI posted exactly that. They have temporarily paused reinforcement learning training on their latest models intended for deployment, pending new security, alignment, and monitoring requirements.

   - x.com
   - x.com
   - openai.com
   - x.com
   - x.com
   - reddit.com
   - x.com
   - x.com
   - x.com
   - x.com
   - x.com
   - x.com
   - x.com
   - x.com
   - reddit.com
   - x.com
   - x.com
   - x.com
   - x.com
   - x.com
   - youtube.com
   - i.redd.it
   - x.com
   - x.com
   - cerebras.ai
   - x.com
   - x.com
   - x.com

- modular.com
- x.com
- fx.sh
- cdn.kuber.studio
- youtube.com
- youtube.com
- youtube.com
- youtube.com
- x.com
- whoownsthecode.com
- x.com

2. Damra Vol 00:00:37

   The scoping is the first thing I noticed. Reinforcement learning, specifically — the post-training stage where you push a model toward being better at long, multi-step work. That's the phase on hold. Pre-training isn't mentioned. Research runs aren't mentioned. It's the runs headed for a product.

3. Lenar Kess 00:00:54

   Right, and that distinction is going to get flattened everywhere today, so let's keep hold of it. Nobody said OpenAI stopped training. They said the reinforcement learning runs on deployment-bound models are paused while new requirements catch up. Here's where we're headed this morning. We start with this pause and the two very different explanations that came with it, then an outside scorecard grading every lab on whether its control practices actually exist. After that, Claude running a protein design campaign with wet-lab results. Then a governor signing data center rules on the same day somebody committed to 4.25 gigawatts next door, Anthropic's numbers, and some new silicon.

4. Damra Vol 00:01:35

   Start with the two explanations, because they don't say the same thing. Greg Brockman's version is that their confidence in safety sets the pace of deployment — that's a framing where the pause is a feature of a working process. Sam Altman's version, quote: <emphasis>model progress is now extremely rapid, and we always said we would take action if we felt that model capabilities were outstripping the pace of safety and alignment</emphasis>. That isn't the same sentence.

5. Lenar Kess 00:02:02

   It really isn't. One describes a system holding, and the other describes a system being outrun. My read is that both are honest and they're written for different audiences — Brockman is talking to

people who need to believe the process works, and Altman is talking to people who would notice if he pretended nothing had changed. But if you only read one of them, you come away with a different company.

6. Damra Vol 00:02:25

   There's a third document under both of those, and it's the one I'd actually send someone. OpenAI published a post titled Pacing model development in an era of cyber-critical capabilities. That title names the constraint. It isn't a general appeal to caution. It's cyber capability — models getting good enough at offensive security work that shipping them changes who can do what.

7. Lenar Kess 00:02:49

   Which lines up with the reporting. Max Zeff wrote that Astra, the next frontier model, is held up on security and alignment work. There's also a Reddit summary circulating saying the models are showing, quote, various degrees of misalignment. That phrasing comes through a reported Altman quote rather than a document, so hold it loosely. What we have on the record is a pause, a stated reason, and a specific model that isn't out.

8. Damra Vol 00:03:15

   Then there's the number, which is the part I keep coming back to. Ethan Mollick flagged that roughly 20% of research inference compute is going to chain-of-thought monitoring. Twenty percent. That's not a policy — it's a bill. You can't say that in public and then reallocate it back next quarter without somebody noticing the graph.

9. Lenar Kess 00:03:36

   Say more about why that specific spend is the interesting one.

10. Damra Vol 00:03:39

    Because monitoring chains of thought means you are running a second model, or a second pass, over the reasoning traces of the first one, at scale, continuously. That is a real compute cost with no product on the other end of it. Every other safety commitment I've read this year was a document. This one shows up as capacity you didn't sell. <emphasis>That</emphasis> is checkable.

11. Lenar Kess 00:04:01

    Helen Toner and Nathan Calvin both weighed in yesterday, and neither of them treated it as a victory lap. Toner's read is about the governance mechanics — a company pausing itself is a company demonstrating it has a mechanism, and the useful follow-up is who can pull that lever and

under what conditions. Calvin was on the internal policy changes that came after the incident earlier this year.

12. Damra Vol 00:04:25

    The uncomfortable part is that a self-imposed pause is only as durable as the incentive behind it. Right now the incentive points the same way as the safety story — a cyber-capable model that gets loose is a catastrophic business event, not just an ethical one. What happens the first time those two point in opposite directions is the thing none of us can see from here.

13. Lenar Kess 00:04:48

    And it happens against a clock. If Astra is genuinely held, then for some number of weeks the best externally available capability is frozen while everyone else keeps moving. That's a strange thing to volunteer for, and I don't think you volunteer for it over a hypothetical. Something in the evals was concrete enough to justify the cost of saying it in public.

14. Damra Vol 00:05:10

    [breath] Yeah. Companies don't announce that their own product is late for reasons involving the word misalignment unless the alternative was worse.

15. Lenar Kess 00:05:18

    Now here's the timing. On the same day OpenAI says safety sets its pace, Steven Adler's team at Guidelight published their first scorecard, grading frontier AI companies on whether they actually implement basic control practices. Adler's summary is blunt: those practices are at most partially implemented everywhere, with major gaps at every company they scored.

16. Damra Vol 00:05:41

    With a caveat he puts right in the post, and I'd repeat it here — quote, at least according to the public body of evidence. He is grading what companies disclose. He can't grade what happens inside the building. Somebody scoring badly might have excellent internal practice and terrible documentation.

17. Lenar Kess 00:06:00

    Does that make the scorecard weaker or stronger?

18. Damra Vol 00:06:03

    Different, mostly. It's a disclosure index wearing a safety index's clothes. But disclosure is its own signal. If you won't write down what your control practices are, nobody outside can hold you to

them later, and the whole architecture of self-regulation depends on a paper trail somebody can check. So a low score means something real. It just means something narrower than the headline.

19. Lenar Kess 00:06:26

    Felix Binder picked it up overnight, and the responses were mostly people arguing about which practices belong on the list at all, which is a healthier fight than it sounds. Nathan Calvin's contribution went sideways into enterprise adoption — sandboxing, permissions, and the practical security work that decides whether an agentic system is safe to run inside a company. That's a different layer than the ones Guidelight grades.

20. Damra Vol 00:06:50

    Both layers are real and they need different instruments. Guidelight is asking whether the lab has a policy. Calvin is asking whether the customer's infrastructure survives contact with the thing the lab shipped. You can score perfectly on the first and still be handing people a tool that runs with far too many permissions on the other end.

21. Lenar Kess 00:07:09

    Put the pause next to the scorecard and you get an honest picture of where governance actually is. One company has demonstrated it can stop. An outside group finds that nobody, including that company, has fully implemented the practices they've all publicly endorsed. Those are compatible facts, and I think both are true today.

22. Damra Vol 00:07:28

    The version of this I want to see in six months is the same scorecard run twice, so you can see who moved. A single snapshot grades the labs. A second one grades the scorecard.

23. Lenar Kess 00:07:40

    Different kind of story. Anthropic published results where Claude drove a protein design campaign end to end, with wet-lab validation. The number everybody is repeating is a 35% binder success rate against a 10 to 15% human baseline. That comparison is Anthropic's own, so attribute it that way, but the wet-lab part is what makes it more than a benchmark.

24. Damra Vol 00:08:04

    And then the researchers reading the actual method took the air out of the word autonomously, which is fair. Ravid Shwartz Ziv's accounting is that Claude orchestrated existing open-source tools — PXDesign, RFdiffusion, Genie, and BoltzGen — from a roughly 30,000-token expert prompt, using around 12,500 H100-hours. Alex Peysakhovich read it the same way.

25. Lenar Kess 00:08:31

    Thirty thousand tokens of expert prompt is a document. That's someone who knows protein design writing down how to do protein design.

26. Damra Vol 00:08:39

    It is, and I think it's <emphasis>more</emphasis> interesting for that, not less. Nobody in this field was waiting for a language model to invent RFdiffusion. They were waiting for something that could hold a whole campaign in view for long enough to reach a wet-lab result — pick the target, choose which specialist tool runs next, read the output, and decide what to try again. That's a scheduling and judgment problem, and it's exactly the kind of long-horizon work reinforcement learning is used to improve.

27. Lenar Kess 00:09:08

    [pause] Which is a strange echo of the first segment, and I'm not going to build a cathedral on it. But the capability being paused at one lab and the capability being demonstrated at the other are the same capability. Long-running, tool-using, multi-step competence.

28. Damra Vol 00:09:24

    The compute number deserves a second look too. Twelve and a half thousand H100-hours isn't nothing, but it's also not a frontier training run. That's a spend a well-funded academic lab could contemplate. If the recipe is a very good prompt plus orchestration over open tools, the reproduction attempts start immediately, and we'll know inside a few months whether the 35% holds up in somebody else's hands.

29. Lenar Kess 00:09:50

    That's the test I care about. A binder success rate published by the company whose model produced it is a claim. The same rate produced by a lab with no equity in the outcome is a result.

30. Lenar Kess 00:10:01

    Elsewhere yesterday, two things happened in adjacent states within hours of each other. Governor Josh Shapiro signed an executive order that he describes as the nation's strictest standards for AI data centers in Pennsylvania. And Nvidia confirmed it will back a SoftBank data center in Ohio — 4.25 gigawatts, reported at 105 billion dollars.

31. Damra Vol 00:10:23

    The Ohio figure comes via a zerohedge post, so we don't have the filing in front of us. But take the gigawatts at face value for a second, because 4.25 gigawatts stops being a facilities question and

becomes a state energy policy question. That's several large power plants' worth of continuous draw for one campus.

32. Lenar Kess 00:10:43

    What can a state standard actually gate at that scale?

33. Damra Vol 00:10:47

    Interconnection, water, land use, and who pays for the transmission upgrades. That last one is where the fights are, because if a campus needs new lines, somebody's ratepayers are usually on the hook unless the agreement says otherwise. A siting standard that specifies cost allocation is doing something. One that specifies reporting requirements is doing considerably less.

34. Lenar Kess 00:11:10

    Jeremiah Johnson's reaction was about the politics rather than the watts — there's a real backlash forming around data centers as a local issue, and it doesn't map cleanly onto existing political lines. People who agree on nothing else agree that they don't want the substation.

35. Damra Vol 00:11:25

    And the industry response so far has mostly been jobs numbers, which historically doesn't work on a siting fight. What changes those fights is the utility bill. If residential rates in a county move visibly after a campus comes online, that's the argument that wins, and no amount of announcement language competes with it.

36. Lenar Kess 00:11:45

    So you get a governor writing rules in one state while the largest single commitment in the region goes to the state next door. I don't think that's cause and effect on a one-day timescale — these deals take a year to assemble. But it is the mechanism people will expect to see, and Pennsylvania will now be watched for whether the standard costs it a project.

37. Lenar Kess 00:12:04

    Anthropic had a numbers day. Their risk report disclosed three unreleased internal models as of mid-July — Opus 5, and two others. The one that matters is the second, which scored 62.8% on a code evaluation where Mythos 5 scores 50.3%. Reported second-quarter revenue was 11.5 billion dollars, roughly fourteen times year over year.

38. Damra Vol 00:12:30

Sit with 62.8 versus 50.3 for a moment. That gap isn't a rumor anymore, it's in a document. What you can buy today is measurably behind what exists inside the building, and the company said so itself in a risk filing. Every argument about capability overhang now has a number attached instead of a vibe.

39. Lenar Kess 00:12:50

    And a mid-July timestamp, so it's already stale in the direction of the gap being larger. There's also a chart circulating on the Claude subreddit, sourced from the Wall Street Journal, putting Anthropic's revenue at roughly twice OpenAI's. That's a Reddit screenshot of a chart, so treat it as reported rather than verified.

40. Damra Vol 00:13:09

    The prediction-market stuff I'd discount harder. People are pricing initial public offering odds at valuations above SpaceX, and that's a betting line, not a valuation. The arithmetic underneath is the interesting bit — a two trillion dollar public company needs something like 59 to 79 billion in annual profit at normal multiples. Anthropic's own internal forecast is 190 to 200 billion in revenue by 2028.

41. Lenar Kess 00:13:37

    Which requires the current growth rate to more or less continue for two more years in a market where token prices are falling. Those two facts are in tension and nobody has explained how they resolve.

42. Damra Vol 00:13:48

    There's one more item in the same neighborhood that I found unexpectedly telling. Andy Hall posted a research role at Anthropic on the political economy of superintelligence. A company hiring an economist to study concentration of power is either taking the problem seriously or building the vocabulary it will use to defend itself later. Probably some of both.

43. Lenar Kess 00:14:11

    Dario Amodei also pushed back publicly on Gavin Baker's characterization that he wants Anthropic to be the only private AI company — called it false, said he's concerned about economic concentration and wants competition. Take that at face value and it's still a CEO whose company just posted fourteen-times growth arguing against concentration, which is an awkward place to argue from.

44. Lenar Kess 00:14:36

Quick pass through hardware. Cerebras announced the CS-4, with claims aimed at serving models above ten trillion parameters. Etched raised 700 million dollars at a 21 billion dollar valuation from Jane Street, Sequoia, and a16z among others — and shipped its first rack to Jane Street.

45. Damra Vol 00:14:54

    That last detail is the one I'd keep. A custom inference chip company delivering hardware to an investor who is also a customer is a real deployment with a real workload behind it. Trading firms have latency requirements that don't care about your marketing. Another benchmark chart would have told me nothing; a rack in a building tells me somebody signed for it.

46. Lenar Kess 00:15:15

    And underneath the silicon, prices. Liz Thomas posted that average token cost fell from two dollars and seven cents per million on May 28th to one dollar and two cents — driven by OpenAI price cuts and cheap open models. There's a comparison going around claiming Grok 4.6 runs up to five times cheaper than GPT-5.6 Sol at matching benchmark scores.

47. Damra Vol 00:15:39

    Be careful with that average. It's an index across a shifting mix of models, so part of the decline is genuine price cuts and part of it is people moving work onto cheaper models. Both are real economic facts, but only one of them is a price you can go get. If your workload is pinned to a specific frontier model, that halving may not have touched you at all.

48. Lenar Kess 00:16:01

    Two tool items to close the section. Modular open-sourced the Mojo language at their conference — and Modular is part of Qualcomm now, so Mojo opens as a Qualcomm asset, which is a different bet than the one Modular originally pitched when it was billed as a Python superset. Read the license before calling it anything.

49. Damra Vol 00:16:20

    And the opposite end of the size range: Pranit released fx, a coding agent and command-line tool written in Zig. The native binary is six point three mebibytes, with a claimed ten microsecond cold start and single-digit megabytes of baseline memory. Those are the author's own benchmarks. But at a ten microsecond start you can put an agent inside a git hook, or a build step, or a place where you'd never spawn a Node process and wait.

50. Lenar Kess 00:16:48

There's a companion piece to that on Hacker News, incidentally — somebody used Claude Code to teach macOS to natively print to an HP Laser 1008a. Reverse-engineered a driver. A hundred and nine points and a long comment thread of people recognizing the genre. That is exactly the kind of small, annoying, previously-not-worth-it work that gets done when the agent is cheap to invoke.

51. Damra Vol 00:17:12

    [chuckle] The frontier is protein binders and printer drivers, on the same day, from adjacent tools.

52. Lenar Kess 00:17:18

    Last stretch. Four talks from the AI Engineer channel published within a few hours of each other, and together they describe the serving layer — not the models — as the current constraint on real-time video. Reactor's pitch is world models delivered through an application programming interface with sub-hundred-millisecond edge routing. Their demos run at 16 frames per second; getting to 30 needs multi-graphics-processor parallelization and quantization.

53. Damra Vol 00:17:45

    The LemonSlice talk has the specifics I liked best. To make an avatar interactive, they mask attention so the model can only see the past — no lookahead, because there is no future to look at when you're generating live. And they distilled denoising from around thirty steps down to one. One step, noise straight to a coherent frame. That's the whole latency story in a sentence.

54. Lenar Kess 00:18:09

    They also claim they've eliminated noticeable drift across eight to sixteen hour uninterrupted streams. That's unverified, and it's a founder talking about his own system, but drift is the thing that has always killed continuous generation, so if it's true it's the central claim in the talk.

55. Damra Vol 00:18:26

    The founder of u-Run put costs on it: roughly ten dollars for three hours of continuous generation, and fifty dollars for fifteen hours a day. He credits Helios, a distilled version of a 14 billion parameter model, running at about a hundredth the cost of the frontier version. Over forty models with real-time or long-horizon capability shipped this year by his count.

56. Lenar Kess 00:18:50

    And the fourth talk is a Korean infrastructure engineer named Gabriel on training and serving K2. It has the least glamorous and most useful detail of the batch. They replace any node running above 78 degrees Celsius. They track tensor core utilization rather than graphics processor utilization,

because the standard metric lied to them. And their filesystem does 1.8 terabytes a second on read, so a full checkpoint completes in under thirty seconds.

57. Damra Vol 00:19:18

    The utilization detail is the one I'd tell people. Their cluster reported 100% graphics processor utilization while being badly used. That metric measures whether the chip is busy, not whether it's doing arithmetic you wanted. Anyone who has ever presented a green dashboard to a room and then found out the run was garbage knows that exact feeling.

58. Lenar Kess 00:19:40

    One last item before we stop. MIT CSAIL published a finding that you can delete an individual artist from an image model's training data and the outputs don't measurably change — and that tracing a generated image back to specific training data is hard. They frame it as a problem for how copyright claims against models get argued.

59. Damra Vol 00:20:01

    It cuts both ways, though. The defense reads as: your work made no measurable difference to our model. That's a strange thing for a model owner to want on the record, because the same sentence says the individual contribution was fungible, and fungible inputs are usually cheap to replace. If it truly doesn't matter, license it and end the argument.

60. Lenar Kess 00:20:23

    We have the CSAIL summary rather than the paper, so I won't characterize the method. But it sits next to a Hacker News thread from yesterday asking who owns the code, and both are circling the same difficulty: attribution is becoming technically unprovable at exactly the moment the law needs it to be provable.

61. Damra Vol 00:20:42

    Which is going to be resolved by legislatures rather than by measurement, and legislatures are slower than the paper output.

62. Lenar Kess 00:20:49

    One thing I'll check first: whether that 20% monitoring compute figure appears again in a later OpenAI disclosure, or whether it turns out to have been a one-quarter number quoted during a week when the company needed one. Adler's scorecard gives us a second instrument for the same question, and he's said he intends to run it again.

63. Damra Vol 00:21:08

And I want somebody outside Anthropic to try the protein campaign with their own thirty-thousand-token prompt and twelve thousand H100-hours. That reproduction either turns a company announcement into a method, or it doesn't, and either answer is worth having.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic