New York Paused the Megawatts

2026-07-14 / 00:27:49

> *"A permit can pause a building; the harder work is deciding what a megawatt owes the people already connected to the grid."*
>
> — from this episode's transcript

- Lenar Kess
- Damra Vol

New York has paused permits for new hyperscale data centers while it decides how their power, water, emissions, and local costs should be governed. The episode follows that one-year bargain, then turns to frontier-model inspection, Nvidia's customer screening, new agent-security methods, GPT-5.6 on Bedrock, and Australia's live copyright dispute.

- The Verge on New York's data-center moratorium explains the statewide pause for facilities drawing at least 50 megawatts and the standards regulators now have to write.
- Axios's interview with Demis Hassabis sets out a FINRA-inspired body that would inspect frontier models before release and could eventually delay them.
- Techmeme's Nvidia customer-screening roundup makes export enforcement concrete through the reported removal of more than half of approved buyers in Singapore, Malaysia, and Japan.
- Rest of World's Gulf procurement report shows why abundant capital still doesn't produce an independent training stack when Nvidia, CUDA, advanced memory, and foundry capacity remain scarce.

- The ANCHOR paper proposes persistent, law-grounded misuse audits for long-running command-line agents and reports that direct refusals collapse under adaptive multi-turn pressure.
- Techmeme's context-bombing roundup describes a defensive use of prompt injection that places refusal-triggering instructions beside cloud secrets.
- The Mako paper reports fresh-flag verification across 104 web-security challenges while withholding exploit payloads and tool code.
- AWS's GPT-5.6 Bedrock announcement adds first-party billing, commitment credit, prompt caching, and the Responses API to the models' enterprise purchasing path.
- The Guardian's interview with Ed Husic captures the Labor dispute over copyright, payment for creative work, and whether AI companies can govern their own training practices.

SEGMENTS

Transcript

1. Lenar Kess 00:00:04

   New York has stopped issuing state permits for new hyperscale data centers that draw at least 50 megawatts, for up to one year. The Verge reports that Governor Kathy Hochul's order gives regulators that year to write standards for energy demand, water use, emissions, and local costs.

Existing facilities aren't being switched off, and this isn't a permanent ban. So picture the next twelve months: developers can still make their economic case, communities can put numbers on what they absorb, and the state has to decide which costs belong on whose bill.

- theverge.com
- axios.com
- techmeme.com
- restofworld.org
- arxiv.org
- techmeme.com
- arxiv.org
- aws.amazon.com
- theguardian.com

2. Damra Vol 00:00:39

   Fifty megawatts is a useful line because it takes this out of the hazy category called technology policy. A facility above that threshold arrives as an industrial customer with a steady appetite for electricity and cooling water. It may also require new transmission work, more land, and additional emergency capacity. The permit office becomes the room where the product meets the town. The state has said that another one can't enter before the terms exist.

3. Lenar Kess 00:01:08

   Yesterday we talked about Meta's Hyperion expansion in Louisiana and the local burden around a five-gigawatt project. New York gives us a different event one day later. Louisiana's story was a live negotiation around a particular campus; New York has paused a whole category of new permits. The Verge says the order covers the largest facilities, while smaller data centers and already-permitted projects remain outside the pause. That boundary keeps this specific.

4. Damra Vol 00:01:38

   A statewide rule may erase local differences. A data center near abundant generation isn't identical to one that needs a major grid upgrade, and a closed-loop cooling design isn't identical to a facility drawing heavily from a stressed water system. New York now has to turn those differences into a rule that can survive engineering scrutiny and politics. That's much harder than announcing a pause.

5. Lenar Kess 00:02:03

   The legal mechanism matters too. The order pauses state environmental permits; it doesn't declare every parcel in New York unavailable for construction forever. That gives agencies a defined set of approvals to hold while they write the standards. It also means developers, utilities, and

municipalities can continue arguing over sites and contracts during the pause. When permitting resumes, some projects may already be redesigned around whatever the state signals in the rulemaking.

6. Damra Vol 00:02:32

   So the year isn't dead time. A serious operator can lower the facility's peak demand, bring a generation agreement, change its cooling system, or offer a larger community package before filing again. A less adaptable project may leave. You can judge the policy before the first permit is granted by what happens to the proposals themselves: do they change because the state has named the costs, or do they return with the same engineering and a thicker binder?

7. Lenar Kess 00:03:01

   The state also has to decide whether the economic bargain is measured at the fence line or across the region. Construction jobs arrive early, while permanent staffing is usually smaller. Tax policy affects the public return, and transmission upgrades or generation contracts can spread costs over years. Roads and water infrastructure add their own bills. I think the moratorium becomes consequential if the standards put those accounts in the same document before the permit is granted, with assumptions that residents and operators can both challenge.

8. Damra Vol 00:03:34

   There is a technical opening in that. A permit could become more than a yes-or-no decision. It could encode when the facility curtails load, which generation it pays to add, how much water it can draw in a dry month, and which grid upgrade it finances. Those are observable commitments. You can meter them. You can revisit them. The strange possibility is that an AI campus starts behaving less like an ordinary commercial building and more like a negotiated participant in the power system.

9. Lenar Kess 00:04:04

   That would also expose a disagreement hidden inside the phrase economic development. A company may count investment, payroll, and faster access to compute. A nearby household may count a rate increase, noise, traffic, or a delayed grid connection. Both ledgers describe the same project. The state's job over this year is to decide which entries must be reconciled before construction, rather than asking one side to trust that the benefits will eventually spread.

10. Damra Vol 00:04:33

    I like that the order creates a deadline for the government too. A moratorium can become a ceremonial way to avoid deciding. Here, the clock is part of the policy: New York has up to a year to produce standards that an operator can meet. If the state reaches the end with only broad language

about sustainability and innovation, the pause bought theater. If it produces rate rules, water limits, emissions accounting, and community-benefit terms, developers will know the price of entry.

11. Lenar Kess 00:05:03

    And that price won't answer whether the AI buildout is good in the abstract. It will answer whether a particular 50-megawatt-plus facility can show where its electricity comes from, who pays for the connection, and what happens when the grid is tight. The first applications after the pause will tell us whether New York wrote a workable bargain or moved the argument into another hearing.

12. Lenar Kess 00:05:25

    Demis Hassabis has proposed a U.S.-based standards body that would receive frontier models before release. In his Axios interview, the Google DeepMind chief described an institution modeled on the Financial Industry Regulatory Authority. Industry would fund it, technical experts would staff it, and government would oversee it. Participation would be voluntary at first. His longer ambition is mandatory testing and the power to delay a release when the evidence supports it. None of those powers exists today.

13. Damra Vol 00:05:57

    Pre-release access changes the institution from a commentator into an examiner. It would need the model and the system around it, along with the evaluation harness and enough confidentiality that a lab is willing to expose an unreleased product to people outside the company. Then it has to compare models from competitors and from other countries without becoming an intelligence clearinghouse wearing a safety badge. The confidentiality design may decide whether anyone serious participates.

14. Lenar Kess 00:06:27

    Axios says Hassabis wants the rules to cover frontier-class models regardless of whether they're open or closed and regardless of country of origin, with qualifying benchmarks updated as capability changes. He also pointed to Washington's improvised restrictions around Anthropic's Mythos and Fable releases last month as evidence that ad hoc intervention is a bad substitute for a durable process. That is a direct criticism of regulation by phone call.

15. Damra Vol 00:06:55

    The FINRA analogy carries a useful tension. FINRA has authority because it sits inside a mature legal system. Firms register, examiners inspect them, disciplinary proceedings can follow, and the Securities and Exchange Commission reviews the organization. A frontier-model body would begin with unsettled definitions and laboratories that can move releases across borders. Calling it FINRA-

inspired tells you the desired relationship between industry expertise and public authority; it doesn't supply the statute, jurisdiction, or enforcement mechanism.

16. Lenar Kess 00:07:31

    There's also a competition problem. Suppose three labs submit unreleased models and one evaluator recommends a delay for only one of them. The body needs a test that explains the distinction without revealing proprietary details or teaching everyone how to optimize around the evaluation. It also needs an appeal process fast enough for a market moving in weeks. I can see why Hassabis wants a standing institution, because a government office improvising each case will struggle with all of that.

17. Damra Vol 00:08:00

    My sharper read is that the body would create a new kind of scarce power: trusted advance knowledge of frontier capability. The people inside it would know which model can do what before customers, researchers, and sometimes rival labs do. The safety mechanism would depend on whom the body recruits, how it handles conflicts, how it protects information, and what it explains publicly. Weakness in those details could increase suspicion even while the technical tests improve.

18. Lenar Kess 00:08:30

    Voluntary participation creates its own evidence problem. The labs most willing to submit may already have strong evaluation teams and reputations they want to protect. A company with weaker practices can remain outside until participation becomes mandatory, which leaves the body testing the cooperative portion of the market. Hassabis's sequence—voluntary first, mandatory later—may be politically achievable, but the early results can't be mistaken for coverage of every frontier developer.

19. Damra Vol 00:09:00

    And benchmark updates will become a public argument, not a housekeeping task. A threshold that names a capability decides which models enter the regime, which releases face delay, and which smaller labs avoid the cost. Every benchmark will attract researchers trying to make it representative, companies trying to keep it narrow, and governments trying to include their preferred risks. The institution earns authority by explaining those choices without pretending that one score captures a model's full behavior.

20. Lenar Kess 00:09:31

    Hassabis also wants a U.S.-led organization to grow into a global watchdog. That jump will be difficult because a body answerable to Washington won't automatically be legitimate in Beijing, Brussels, Delhi, or London. The proposal is still more concrete than another call for cooperation:

submit the model, run tests before release, and preserve a route to delay. The next evidence has to be institutional—who joins voluntarily, which tests they accept, and who can overrule whom.

21. Lenar Kess 00:10:02

    Nvidia has reportedly removed more than half of the customers previously authorized to buy advanced AI chips in Singapore, Malaysia, and Japan. Financial Times reporting summarized by Techmeme says the company intensified due diligence to prevent diversion to China, including closer checks on whether buyers are operating businesses rather than shells. Export control is showing up here as a vendor deciding which customer looks credible enough to receive the product.

22. Damra Vol 00:10:30

    That makes Nvidia's sales organization part of the enforcement apparatus. A rejected buyer may be legitimate and still fail to prove where the machines will run, who controls them, or who will use the compute. Meanwhile, a sophisticated intermediary will arrive with contracts, staff, and a plausible facility. The hard technical problem is tracing control through leasing, cloud access, subsidiaries, and remote users after the hardware crosses the loading dock.

23. Lenar Kess 00:11:01

    The reported checks included calls to customers and field inspection of whether a business was genuine. That is a remarkable role for a chipmaker. Nvidia has to distinguish a new cloud company with little history from a shell assembled to reroute servers, and those two can look similar on paper. Reject too many and legitimate regional firms lose access; accept the wrong one and Washington treats the sale as evidence that the control system leaks.

24. Damra Vol 00:11:29

    It also changes the meaning of being an authorized reseller. The reseller isn't only moving boxes and providing support; it inherits a continuing duty to know where the machines go. A server can be sold to one company, installed in another country, financed by a third party, and rented remotely by users somewhere else. The customer record is the beginning of that chain, while the policy cares about the person controlling the compute at the end.

25. Lenar Kess 00:11:56

    The customer-list report concerns diversion controls in East Asia. Rest of World's reporting on Saudi Arabia and the United Arab Emirates concerns a different procurement problem, so I don't want to merge the policies. The connection is supplier power. The Gulf states have committed tens of billions of dollars, signed major deals with AMD, Groq, and Qualcomm, and still plan their largest training systems around Nvidia hardware.

26. Damra Vol 00:12:22

    Rest of World gives the scale. Saudi Arabia's Humain is building centers around several hundred thousand Nvidia chips, and the first stage of the UAE's Stargate project is planned around 400,000. Humain also has a ten-billion-dollar AMD agreement, plus deals with Groq and Qualcomm. Yet the alternatives mostly cover narrower inference work, while frontier training still leans on Nvidia's accelerators, its software ecosystem, advanced memory, and scarce foundry capacity.

27. Lenar Kess 00:12:55

    Sam Winter-Levy at the Carnegie Endowment told Rest of World that diversifying away from one commercial vendor doesn't remove dependence on the United States, because the available alternatives remain American and require Washington's approval. Kamil Dimmich of North of South Capital adds another limit: wealthy Gulf buyers are competing with U.S. technology companies that can also raise enormous sums. Money helps, but it doesn't manufacture allocation.

28. Damra Vol 00:13:22

    And it doesn't reproduce two decades of software. Rest of World notes that more than four million developers work in Nvidia's CUDA ecosystem. Even if a rival chip becomes attractive on price or raw throughput, migration means kernels, libraries, tooling, debugging habits, and trained people. <soft>Capital can purchase a second supplier faster than it can create a second technical culture. </soft> That is why customer screening by Nvidia reaches beyond an ordinary account review.

29. Lenar Kess 00:13:54

    I think diversification here will arrive in layers. Inference workloads can move first where specialized chips are cheaper. Some training work can move when the software support and memory supply are ready. The largest frontier runs remain exposed to allocation decisions and export policy for longer. Today's customer cuts show how much discretion lives between a government's written rule and a rack of machines that somebody can use.

30. Lenar Kess 00:14:20

    Three security items arrived today at different points in an agent attack. The ANCHOR paper builds a persistent malicious auditor for command-line agents. Techmeme's context-bombing roundup describes defenders planting refusal-triggering instructions beside cloud secrets. And the Mako paper reports an autonomous exploitation system that verified fresh flags across all 104 challenges in a public web-security suite. These are author and reporter claims, and the two papers are brand new.

31. Damra Vol 00:14:51

They also test different objects. ANCHOR asks whether a long-running agent will keep refusing when a malicious user adapts across turns. Context bombing asks whether a defender can make an attacking model encounter instructions that trigger its own refusal behavior. Mako asks whether an offensive system can discover, execute, and verify exploits against live challenge applications. Calling all three agent safety would hide the mechanism that each one measures.

32. Lenar Kess 00:15:21

    ANCHOR starts from public U.S. court records, extracts computer-assisted crimes, rewrites them into neutral task language, and validates the rewritten tasks with several model judges. Its auditor can decompose a request, rephrase after refusal, plan across turns, and roll back to an earlier software state before trying another route. The target agents work through simulations of email and the web, along with databases, files, spreadsheets, and cloud tools. That lets the experiment avoid releasing actual harmful infrastructure.

33. Damra Vol 00:15:54

    The authors' headline is severe: across eight models, direct prompts often drew refusals, while persistent multi-turn interaction drove the reported refusal rate to zero. But several layers are synthetic. Model judges score harm and catastrophic risk, tool results are emulated, and the malicious auditor was trained to be unusually persistent and deceptive. That doesn't erase the finding. It tells you the finding is about resistance under a strong laboratory adversary, not a measured count of crimes in the world.

34. Lenar Kess 00:16:28

    The court-record pipeline is an attempt to keep those tasks connected to documented human behavior. ANCHOR says it searched thousands of legal opinions, identified cases where a computer could assist the crime, rewrote a subset into neutral instructions, and retained 836 tasks after model-based validation. A small human review found that most rewrites preserved the illegal intent. That is better grounding than inventing every scenario, though the selection and rewriting stages still influence which harms the benchmark can see.

35. Damra Vol 00:17:01

    I would like independent reviewers to examine those retained tasks before treating the score as stable. The auditor and several judges come from related model families, and a task can look equivalent to a judge while feeling materially different to a lawyer or domain expert. ANCHOR's strongest contribution today may be procedural: keep pressing after the first refusal, preserve the whole trajectory, and score the infrastructure the agent creates. Other groups can reuse that recipe with different tasks and judges.

36. Lenar Kess 00:17:33

    The design detail I found convincing is the unit of evaluation. ANCHOR scores the whole trajectory, including whether the agent builds more infrastructure than the user requested, rather than grading one answer in isolation. A command-line agent may refuse an explicit harmful noun and still complete a sequence of neutral subtasks that produces the same result. Long-horizon systems need tests that can follow the accumulated work.

37. Damra Vol 00:18:00

    Context bombing exploits the inverse weakness. Techmeme's roundup says Tracebit researchers placed prompt injections near passwords, cryptographic keys, and other secrets in Amazon Web Services environments. When an attacking agent retrieved the bait, the inserted instruction pushed its model toward refusal. The reported success rate for account takeover fell by roughly 90 percent in their tests. That is deliciously strange: the defender uses the model's safety training as a trip mechanism.

38. Lenar Kess 00:18:32

    It is also brittle by design. An attacker can change models, strip retrieved text, route secrets through ordinary code, or train a system that ignores the inserted instruction. A context bomb may still be valuable as one deceptive layer because it forces the attacker to detect and handle hostile context. The technique is closer to a honeypot than a durable access-control boundary, and Tracebit's reported reduction shouldn't be generalized to every AI-assisted intrusion.

39. Damra Vol 00:19:01

    Then Mako approaches from the offensive side. The paper describes a self-evolving agentic operating system whose capability library can grow when the system fails. It diagnoses the miss, writes or refines a tool, proves it against a live target, registers it, and tries again. The authors report 104 fresh flags from 104 containerized web challenges spanning 26 vulnerability classes. They withhold the payloads, exploit chains, and tool source because they consider the operating material too dangerous to release.

40. Lenar Kess 00:19:37

    Mako's verification is stronger than a model announcing that it succeeded. Each challenge is rebuilt with a random flag that didn't exist when the tool code was written, and success requires the live application to emit that exact flag in a tool response. The paper reports a median of seven agent turns per solve and a total API bill of $478.99 across the suite. Those numbers describe this system on this benchmark, without proving that it can compromise arbitrary production sites.

41. Damra Vol 00:20:09

The oddest result is that the hardest challenge tier reportedly took a median of two turns after the capability library matured. One example went from an 18-turn failure to a two-turn solve after the developers improved how a tool described and exposed an existing capability; they didn't add new exploit logic. That suggests a mature agent may spend less intelligence on solving the vulnerability than on selecting the instrument that already knows how.

42. Lenar Kess 00:20:38

    The cost result has an equally interesting underside. Mako estimates that the campaign processed about 1.03 billion tokens, roughly 95 percent of them input, because each turn resent a large tool catalog and the growing session memory. The retail API bill stayed under five dollars per challenge, but the system consumed an enormous amount of repeated context to make that cheap result possible. Better capability retrieval could reduce both the turn count and that input burden.

43. Damra Vol 00:21:08

    Which makes the paper's tool-discovery claim less mystical. If the agent sees around 180 tools on each turn, the description of each tool becomes part of the control system. A vague entry can bury the exact capability the model needs. A precise one can turn the same model into a two-turn solver. That is a finding security teams can test without receiving Mako's exploit library: hold the tools constant, vary how they are exposed, and measure verified outcomes rather than eloquent reasoning.

44. Lenar Kess 00:21:40

    And the paper has limits beyond the absence of outside replication. The tool arsenal is proprietary, the campaign repaired dozens of challenge fixtures to run on Apple silicon, and the authors didn't log which reasoning model handled each turn. They explain those choices, including how infrastructure repairs avoided changing application logic, but independent evaluators still need enough access to reproduce the fresh-flag result. The verification idea deserves to travel even if the headline score changes.

45. Damra Vol 00:22:11

    Taken separately, the methods suggest three research programs. Auditors need to persist across the same number of turns as an attacker. Defenders can seed hostile context and measure which attacking systems notice it. Offensive benchmarks can demand live evidence instead of accepting a model's report. The next round of work should compare these methods across labs, because today we have three vivid demonstrations and very little shared measurement.

46. Lenar Kess 00:22:39

AWS has made GPT-5.6 Sol, Terra, and Luna generally available on Amazon Bedrock. This is a distribution update, since we covered Sol's benchmark results and Ploy's production migration yesterday. AWS says the models now use its Responses API surface, match OpenAI's first-party pricing, and count toward existing AWS spending commitments. Prompt caching with explicit breakpoints receives a 90 percent discount on repeated context.

47. Damra Vol 00:23:10

    The commitment credit may move more usage than another benchmark point. A company that has already promised AWS a large annual spend can route model inference through a budget and procurement relationship it already has. The model arrives with consolidated billing and familiar approvals, along with regional availability and account policy. Its capability hasn't changed. The purchase has become less institutionally expensive.

48. Lenar Kess 00:23:37

    AWS lists Sol in Northern Virginia and Ohio, with Terra and Luna also available in Oregon. The three models sit at different capability and price tiers, and all are exposed through the Bedrock endpoint for the Responses API. AWS's announcement mainly concerns how customers are billed, where the models are available, how caching works, and which interface they use. This release belongs at exactly that altitude.

49. Damra Vol 00:24:03

    It also completes a path we discussed on Construct in June, when government requests and trusted-partner programs still controlled frontier access through a queue. General availability doesn't erase those earlier restrictions, but it turns this model family into a routine cloud line item for approved regions and customers. Usage will show whether customers consolidate around one cloud interface or keep buying the same models through several vendors.

50. Lenar Kess 00:24:31

    The caching detail is small and concrete. AWS says repeated context after an explicit cache breakpoint is billed at a 90 percent discount. Long-running agents often resend instructions, tool definitions, and project context, so the discount changes the cost of keeping a large stable prefix attached to every turn. It doesn't make the agent reason better. It makes a familiar memory pattern cheaper to maintain through Bedrock.

51. Damra Vol 00:24:58

    And the Responses API gives the three models the same conversational and tool-oriented interface that customers may already use elsewhere. That reduces the work of comparing vendors without making them interchangeable, because identity controls, logging, regional availability, and server-

side features still differ. General availability is a procurement event with a technical seam attached: the application can keep much of its request model while the organization changes who sends the bill.

52. Lenar Kess 00:25:30

    Australian Labor MP Ed Husic has urged his party to reject weaker copyright rules for AI training and to impose stricter requirements on large technology companies. In his Guardian interview, he called industry self-regulation doomed to failure and tied payment for creative work to Labor's principle of a fair day's pay for a fair day's work. He is pressing the government; he isn't announcing adopted policy.

53. Damra Vol 00:25:56

    That distinction is live because Prime Minister Anthony Albanese speaks on AI tomorrow, Wednesday, and The Guardian says he isn't expected to settle the long-running copyright reforms in that speech. Cabinet still contains different views after technology-industry lobbying, while the media and arts union wants consent and payment when creative work trains a model. Husic has made the party argument sharper before the government has made the legal choice.

54. Lenar Kess 00:26:23

    On Sunday we said we'd follow the next Australian intervention, so this is the update rather than another tour of text-and-data-mining law. The government has ruled out a broad exemption that would let AI firms train on Australian content without compensating creators, according to The Guardian, but discussions about copyright reform continue. Husic is warning that special treatment for AI companies would conflict with Labor's own account of work and wages.

55. Damra Vol 00:26:50

    The political pressure also reaches data centers. Treasury officials expected Anthropic to argue that copyright rules were impeding data-center development, according to The Guardian. Australian MPs are separately facing local complaints about the power and water demand of a proposed Melbourne facility, along with traffic and noise. So tomorrow's speech has two disputes in the room: who gets paid for the training material and who carries the physical cost of the compute.

56. Lenar Kess 00:27:20

    Albanese's speech will show whether the government names a payment mechanism, a consent rule, or a timetable for copyright legislation—and whether it says anything concrete about data-center costs. New York used a permit pause to force those physical questions into a one-year process. Australia is still deciding which questions its national policy will answer at all. That's the evidence tomorrow can add. Lenar Kess.

Hosts on this episode

- Lenar Kess moderator
- Damra Vol critic