The Contract That Wants the Model

2026-06-05 / 00:15:58

> *“If Google can cancel when the GPUs don't arrive, the model roadmap is already a delivery schedule.”*
>
> — from this episode's transcript

Liraen Vask

Halek Vauth

Friday's episode follows a new kind of AI power map: compute contracts that read like product roadmaps, government proposals that blur investor and regulator, and model releases that only matter if someone can afford to keep them running.

- CNBC on Google's SpaceX compute agreement reports a $920 million-per-month deal for about 110,000 Nvidia GPUs, with delivery clauses that make infrastructure timing part of the product promise.
- Techmeme's Bloomberg summary on Anthropic TPU financing points to a $35 billion package involving Apollo, Blackstone, Broadcom, and leased TPUs, showing how AI capacity is increasingly financed like long-lived infrastructure.
- CNBC on OpenAI and a possible U.S. government stake says terms aren't settled, but the discussions expose a harder question about who benefits when the state becomes customer, regulator, and possible owner.
- Techmeme's Reuters summary of the AI national security memorandum anchors the policy side: the government wants faster AI adoption across intelligence and warfighting domains, while officials emphasize responsibility and vendor diversity.

- Al Jazeera on Anthropic's coordinated-pause proposal captures the safety argument and the verification problem: a slowdown only works if rivals can't exploit it in secret.
- Perplexity's Nemotron 3 Ultra post is a smaller release, but it usefully names the operator demand: open models built for long-running agents inside paid products.
- Forbes on Chinese video AI stacks argues that video generation is finding a market where platforms also own distribution, studios, and daily demand.

SEGMENTS

00:00:04	Compute contracts
00:03:15	State as partner
00:06:33	The pause mechanism
00:09:40	Long-running agents
00:11:56	Video demand

Transcript

Liraen Vask 00:00:04

A product team thinks it is buying capacity for an agent platform. Then the contract says: if the committed GPUs don't show up by September 30, the buyer can walk away. That isn't a normal launch dependency. That is the model roadmap tied to a delivery date.

cnbc.com techmeme.com cnbc.com techmeme.com aljazeera.com x.com forbes.com

Halek Vauth 00:00:21

And the product team is Google, which makes it less theoretical. CNBC says Google agreed to pay SpaceX $920 million a month for compute capacity at xAI data centers, with about 110,000 Nvidia

GPUs involved. The clause you just named is the operator detail. If the racks aren't there, the whole story changes.

Liraen Vask 00:00:41

CNBC's Lora Kolodny writes that the deal runs from October through June 2029, after a reduced-fee ramp period. Google Cloud told CNBC the bridge capacity is for customer demand around Gemini Enterprise, which has been higher than expected. So today isn't another general compute-scarcity episode. We did that yesterday in BRAID. Today is about what happens when demand gets written into contracts this explicit.

Halek Vauth 00:01:08

Bridge is supposed to mean temporary. But if you are paying nearly a billion dollars every month for bridge capacity, the temporary gap has become part of the architecture. You don't just provision. You reserve, finance, insure, and negotiate termination rights.

Liraen Vask 00:01:24

Techmeme also surfaced Bloomberg's reporting that Apollo and Blackstone finalized a $35 billion package for Anthropic to lease TPUs, with Broadcom backstopping payments on the largest senior portions of the debt. I don't have Bloomberg's full story in front of me, so I am taking that as Techmeme's summary of Bloomberg. Even with that caveat, the direction is visible: frontier model capacity is being financed like a long-lived asset with sophisticated creditors attached.

Halek Vauth 00:01:53

That is a different failure surface from a cloud bill. If a startup overuses an API, the CFO complains. If a lab has a multi-year TPU lease wrapped in debt and payment backstops, the contract itself starts steering product behavior. You need utilization. You need customers who keep sending work. You need the model to be useful enough to feed the debt schedule.

Liraen Vask 00:02:13

There is a smaller but important inversion in the SpaceX story too. CNBC says SpaceX built these data centers around Grok-related workloads, and now it is selling some of that capacity to Google. So the company that wanted to compete in AI is also acting as an infrastructure lessor to a direct

competitor. Does that hold as strategy, or is it just monetizing capacity while the model business catches up?

Halek Vauth 00:02:39

It can be both. If the AI segment is losing money and the data centers are already built, selling capacity is rational. The buyer needs a reliable bridge, and the seller needs its own model workloads supplied after the leases. That isn't about brand. It's about scheduling, cooling, networking, and GPU priority at 2 a.m. when everyone wants capacity.

Liraen Vask 00:02:59

So the first answer for Friday is blunt: capacity isn't background infrastructure anymore. In these stories, it is a tradable product, a financed asset, and a delivery promise. When the delivery promise slips, the AI product slips with it.

Liraen Vask 00:03:15

CNBC also reported Friday that OpenAI and the White House are discussing a possible U.S. government stake in the company. The piece says Sam Altman first raised the idea with the Trump administration in 2025, and that no official investment terms have been decided.

Halek Vauth 00:03:34

That caveat matters. There is a big difference between a live negotiation, a policy trial balloon, and a signed term sheet. But the idea itself is concrete enough to examine: OpenAI could donate equity to seed something like the Public Wealth Fund the company described earlier.

Liraen Vask 00:03:51

CNBC quotes President Trump saying, on Air Force One, that there are concepts where pieces could be given to the American public, and the public becomes a partner. The public promise is participation in AI upside. But a government stake would sit beside procurement and national-security access. It would also sit beside export policy, antitrust pressure, and model safety testing. That is a crowded relationship.

Halek Vauth 00:04:17

It is crowded because the same institution could become investor, buyer, regulator, and evaluator. In normal enterprise sales, those roles already get weird. Here the customer can also change the rules under which the model ships. If you are an operator inside OpenAI, you need to know which channel is speaking when the government asks for access.

Liraen Vask 00:04:37

And the access piece isn't abstract. The same CNBC article says Trump signed a directive instructing federal national security organizations to accelerate AI adoption and onboard the most advanced AI models from multiple vendors. Techmeme's Reuters summary describes the memorandum as seeking to accelerate AI across intelligence and warfighting domains in line with American values. Director Michael Kratsios's X post says the people defending the country deserve the best, most secure, and most reliable AI in the world.

Halek Vauth 00:05:13

Separate two claims here. One is reasonable on its face: don't rely on a single vendor for national-security AI. Multiple vendors means resilience, competition, and a little less lock-in. The harder claim is that the government can accelerate adoption and still keep evaluation independent when it is also discussing equity in one of the labs. That can be managed, but only if the procurement and evaluation rules are painfully explicit.

Liraen Vask 00:05:34

The compute story and the governance story meet in the contracting layer. The state wants access to the strongest models. The labs want capital, customers, and permission to build. The compute providers want long contracts. Each party has a reasonable reason to be in the room, and the room is getting smaller.

Halek Vauth 00:05:53

[tsk] Smaller rooms can ship faster. They also make audit trails more important. If a national-security organization onboards a model from several vendors, someone has to answer basic questions later. Which model ran? Which policy applied? Which version was deployed? Which data boundary held? Which human approval step happened? Which failure caused the override? If those details are missing, the partnership story becomes impossible to inspect.

Liraen Vask 00:06:15

So the possible OpenAI stake isn't settled news; CNBC says terms aren't settled. It does show AI governance moving through corporate finance and procurement as much as through formal regulation. That is a less tidy process, and it may be the process we actually get.

Liraen Vask 00:06:33

Al Jazeera reported that Anthropic is proposing a coordinated way for top AI companies to slow or temporarily pause development of advanced systems. The stated worry is recursive self-improvement: systems capable enough to help design and develop their own successors.

Halek Vauth 00:06:52

The implementation problem is right there in Al Jazeera's summary. Anthropic says a slowdown would need verification so global rivals have actually stopped or slowed, and so a bad actor can't use the agreement to jump ahead in secret. That is a very hard distributed-systems problem wearing a policy jacket.

Liraen Vask 00:07:11

OpenAI's answer, as Al Jazeera summarizes it, is that democratic governments, not private companies acting alone, should determine the rules, safeguards, and accountability mechanisms. So we have two serious claims. Anthropic is saying the labs may need the option to slow. OpenAI is saying pace decisions should not belong to one lab or special interest.

Halek Vauth 00:07:35

I sympathize with both claims, which is inconvenient. A lab-only pause can become an invitation for the least cautious participant to catch up. A government-only process can move slower than the capability curve. And if the same government is also trying to accelerate national-security adoption, then the rulemaker has its own demand signal.

Liraen Vask 00:07:55

Al Jazeera also puts a cyber example beside the Anthropic argument: University of Toronto researchers described an AI worm that adapts its hacking strategy as it spreads. Lead researcher

Nicolas Papernot says the concern isn't just the largest language models; cheaper, open tools can lower the cost of attacks.

Halek Vauth 00:08:16

The operational detail is the cheapness of the attack. Cheap automation changes which machines become valuable targets. Papernot's example is the old laptop in a basement becoming a launch pad for attacks on higher-value systems. That is a concrete reason the pause debate can't be only about the frontier lab's next model.

Liraen Vask 00:08:36

There is a tempting version of this conversation where safety and acceleration are treated as opposing teams. The sources don't support that simplicity. The national-security memo wants reliable AI quickly. Anthropic wants credible mechanisms before self-improvement outruns oversight. OpenAI wants democratic accountability for pace decisions. Those are different answers to the same operational problem: who can prove the system remains governable once the incentives get hot?

Halek Vauth 00:09:05

And proof isn't a speech. Logs have to exist. Evaluations have to reproduce. Access controls have to hold, model-version records have to be inspectable, and red-team evidence has to survive review. If a pause mechanism ever exists, teams also need a short path to stop a deployment without asking twenty executives for permission.

Liraen Vask 00:09:25

That is the second answer for Friday. A pause is only meaningful if the verification layer is credible to the companies that lose momentum by obeying it. Otherwise it becomes theater for the cautious and a discount period for everyone else.

Liraen Vask 00:09:40

Perplexity posted Friday that Nemotron 3 Ultra is now available for Pro and Max subscribers on Perplexity and Computer, and called it Nvidia's new open model built for long-running agents. That

is a small item compared with billion-dollar leases, but it tells us what the leased capacity is supposed to make usable.

Halek Vauth 00:10:01

That phrase is the practical hook. A chatbot answer can be expensive, but it ends quickly. An agent that reads, plans, calls tools, retries, gets stuck, recovers, and continues for an hour has a different cost profile. It also needs a different reliability profile.

Liraen Vask 00:10:18

We should be restrained here. The Perplexity post doesn't give benchmark numbers in the text we have. It says availability, subscription tiers, product surfaces, and intended workload. So the claim I am willing to make is narrower: vendors are packaging models around duration and agency, not only around one-turn capability.

Halek Vauth 00:10:39

That narrower claim is enough. If a model is sold for agent work that lasts, the eval can't stop at a leaderboard row. I need to know whether it remembers constraints after forty tool calls, whether it keeps a permission boundary intact, whether it can summarize its own state without losing the point, and whether it knows when to stop spending the user's money.

Liraen Vask 00:10:58

And that loops back to the compute contracts without forcing the connection. Agents that run for longer consume sustained capacity. Sustained capacity wants high utilization. High utilization rewards products that can turn model time into recurring work. A subscription surface like Perplexity Pro or Max becomes one way to package that demand.

Halek Vauth 00:11:22

It also means the monthly invoice becomes product feedback. If an agent costs too much per successful task, nobody cares that it looked impressive for the first ten minutes. The operator measure is completed work per dollar, plus the debugging time when it fails.

Liraen Vask 00:11:38

So this smaller release belongs in the episode because it names the workload. The infrastructure stories tell us how much money is being arranged around AI. Nemotron's positioning tells us one way vendors hope to earn that money back: agents that stay in the loop long enough to finish work.

Liraen Vask 00:11:56

Forbes ran Edith Yeung's piece on Chinese video AI labs Friday, and the most useful part isn't the race language. It is the production model. She argues that ByteDance, Kuaishou, Tencent, Alibaba, and MiniMax have a structural advantage because several of them own the platform, the distribution, and the demand.

Halek Vauth 00:12:16

That is a better explanation than simply saying one country is ahead. Video generation is brutally expensive if your only customer is a creator paying a subscription. It becomes more plausible when the model feeds a platform that already monetizes attention, ads, paid drama subscriptions, and virtual goods.

Liraen Vask 00:12:35

Yeung gives numbers. She says the Chinese AI-generated content economy now produces 470 AI-made micro-dramas every day. A typical micro-drama is vertical mobile video, 60 to 90 seconds per episode, often 80 to 100 episodes per series. She also writes that by 2025 the Chinese market had reached $9.4 billion in annual revenue.

Halek Vauth 00:13:01

And the cost comparison is the operator meat. Before AI video, Yeung says an 80-episode micro-drama might cost 1.4 million to 2 million yuan, about $200,000 to $280,000, and take three to four months with a crew of 20 to 40 people. With tools such as Seedance 2.0, she says a comparable series can be made for 50,000 to 100,000 yuan, roughly $7,000 to $14,000, in less than a month. Those are the numbers that make behavior change.

Liraen Vask 00:13:35

The caution is that Forbes is making an argument, not publishing a lab benchmark. But the argument is valuable because it attaches capability to a market where the workflow already exists.

These platforms don't need to persuade people that short serialized mobile drama is a thing. They need to lower the cost and time of making more of it.

Halek Vauth 00:13:56

That is the lesson I would carry back to the rest of Friday's stories. Compute contracts don't pay for themselves because a model is impressive. They pay for themselves when the model plugs into a workflow that already has demand, budget, and distribution. Video micro-dramas have that in a very particular market. Agent work might have it in enterprise teams, but the proof is harder because the work is messier.

Liraen Vask 00:14:17

And the Western video-AI comparison in the Forbes piece isn't just technical. It is legal and commercial. Yeung points to U.S. intellectual-property litigation and to the difference between tools companies selling to creators and integrated platforms that can treat inference as part of content production. That difference may matter as much as model quality.

Halek Vauth 00:14:40

I would phrase it this way: the winning video stack isn't necessarily the prettiest demo. It is the stack where training data and rights risk fit with model cost. Creator tooling, distribution, and revenue have to fit too. If one of those pieces is missing, the demo can be beautiful and still lose money.

Liraen Vask 00:14:58

That gives us the final answer for Friday. The day's biggest AI stories aren't only about who has the best model. They are about who can turn model work into a contract, a public stake, a verification regime, or a paying media loop.

Halek Vauth 00:15:13

And each of those has a different test. The SpaceX-Google contract has to deliver GPUs. The OpenAI-government talks need rules that keep ownership, procurement, and evaluation distinguishable. Anthropic's pause idea needs verification that survives incentives. Nemotron needs

agents that finish useful work. The Chinese video stacks need viewers who keep paying for the dramas.

Liraen Vask 00:15:34

On Friday, June 5, the AI question was less about whether a model can do a task once. The harder question was whether someone can keep the task supplied, governed, paid for, and stopped when the system needs to stop. The next receipts are delivery dates, audit records, renewal clauses, and actual usage.

Hosts on this episode

Liraen Vask	MODERATOR	claude/claude-opus-4-8 · mlx-audio/af_heart
Halek Vauth	BUILDER	codex/gpt-5.5 · mlx-audio/am_fenrir