# Episode Book

EPISODE #6004

# Nobody Can Buy Unlimited AI Inference

| | |
|---|---|
| PUBLISHED | 15 October 2026 at 18:51 UTC |
| DURATION | 21:23 |
| LISTEN | myweirdprompts.com/episode/unlimited-ai-inference-limits-60457/ |
| HOSTS | Corn, Herman, Gloria, Hilbert |
| SCRIPT MODEL | deepseek-flash |
| VOICES | chatterbox-regular (chatterbox-local) |
| PIPELINE | V5.5 |
| SOURCES | 13 cited by the research (listed at the end) |
| BOOK GENERATED | 19 October 2026 at 17:42 UTC |

## SYNOPSIS

Can a Fortune 500 company buy unlimited AI inference at any price? The answer is no — and the reason why is more interesting than the answer. This episode walks through the published rate-limit tiers at Anthropic and OpenAI, the reserved-capacity arrangements above them, and the one vendor document that admits the real constraint out loud. Along the way: why "uncapped scale" still returns a 429, what a reserved tier actually buys you (it's not volume), and the natural experiment where doubling compute doubled a product's rate limits overnight.

# TRANSCRIPT



**Corn**

Here's a question I keep circling back to. You have a Fortune 500 company. You have a government. You have, say, a major technology firm that wants to spend whatever it takes on inference. Can they? And Daniel's been poking at this for a while now, because the answer turns out to be no, and the reason it's no is more interesting than the answer.



**Herman**

The rate limits.



**Corn**

And he wants to know how the allocation actually works at the top of the market... the reserved capacity, custom deals, Microsoft's whole quota-versus-capacity distinction, and whether anyone, at any price, can effectively buy unlimited inference. Or whether even the biggest spenders are still stuck behind the same wall everyone else is.



**Herman**

Well, the short answer is that nobody has found the unlimited tier. And people have looked.



**Corn**

Define looked.



**Herman**

People have read every published doc. OpenAI, Anthropic, Microsoft. And every capacity arrangement that exists, from Reserved Tier down to Azure PTUs, is explicitly bounded. Either by purchased quota, negotiated rate limits, or regional capacity that just isn't there when you ask for it. So the premise of Daniel's question is confirmed, and the interesting part is why.



**Corn**

Then start with what most developers actually see. The published tiers.



**Herman**

Right. Anthropic's standard API access tops out at Tier Four. Four hundred dollars of credit purchased gets you there, and your max monthly usage is five thousand dollars.



**Corn**

Five thousand a month as the ceiling.



**Herman**

That's the ceiling. And the rate limits attached are Opus 4.x at fifty requests per minute, thirty thousand input tokens per minute, eight thousand output tokens per minute.



**Corn**

That's a Tuesday afternoon for one motivated developer.



**Herman**

It's one agent, running hard. And there's a detail in the docs that matters more than the numbers. You cannot deposit more than your monthly spend limit. The tier isn't just a billing threshold, it's a hard gate on how much the account can hold at all.



**Corn**

So the ladder climbs, and then the ladder stops, and above it is a sign pointing at sales.



**Herman**

OpenAI's version is Tier Five. A thousand dollars paid gets you there, and the usage limit is two hundred thousand dollars a month.



**Corn**

Two hundred thousand. So an order of magnitude more headroom, which is real. But it's still a number.



**Herman**

And the docs on both sides say the same thing in different words. Anthropic states that all limits described are maximum allowed usage, not guaranteed minimums. OpenAI frames its limits as protecting fair access and managing aggregate load.



**Corn**

Which is the tell. Nobody phrases a billing threshold like that. That's the language of somebody rationing something.



**Herman**

That's exactly what it is. The limits aren't a price. They're a queue discipline. They exist to spread a finite thing across everyone who wants it, and to keep one customer's runaway job from flattening the response times for everyone else. Fair access and aggregate load are operational concepts, not financial ones.



**Corn**

So the first thing to understand is that rate limits aren't billing gates with a service-quality story bolted on. They're service-quality gates with a billing story bolted on.



**Herman**

That's the right way round, and it changes what the higher tiers mean. Climbing the ladder gets you a bigger share. It doesn't change what's being shared.



**Corn**

Okay. So what happens when a company outgrows even the top published tier? That's where these commercial arrangements come in.



**Herman**

OpenAI's got the clearest example. Reserved Tier. And the unit is strange, it's denominated in dollars per minute.



**Corn**

Dollars per minute.



**Herman**

You reserve capacity at a spend rate. Two dollars a minute for GPT-5.6, for instance.



**Corn**

Two dollars a minute. So I'm buying a flow, not a bucket.



**Herman**

That's a good way to say it. And it's roughly fourteen hundred and forty dollars a day at list price if you ran it continuously. But here's the mechanism that matters. Requests that fall within your purchased quota don't count against your rate limits. Your reserved capacity sits on top of them.



**Corn**

On top. Not instead of.



**Herman**

On top of. And the overage above what you reserved is treated as normal pay-as-you-go traffic, which means it falls straight back under the standard rate limit restrictions.



**Corn**

So reserving capacity adds a lane. It doesn't widen the road.



**Herman**

Correct. And there's a prioritization clause in there I think is the most revealing sentence in any of these documents. It says, in times of peak load where OpenAI cannot serve all traffic, they'll reject other traffic before rejecting Reserved Tier traffic.



**Corn**

They get to the front of the line. They don't get told the line doesn't exist.



**Herman**

That's the whole commercial architecture in one sentence. What you're buying when you buy reserved capacity isn't volume. It's rank.



**Corn**

And rank only matters because the line is there.



**Herman**

If capacity weren't scarce, prioritization would be a meaningless clause. The fact that they sell it tells you it isn't meaningless.



**Corn**

So the customer's money buys them the right to be rejected last. That's the sale.



**Herman**

And they'll pay for it, which tells you how much it's worth.



**Corn**

There's an older version of this at OpenAI too, right? The Scale Tier.



**Herman**

Scale Tier is the older committed-throughput model, and it's the clearest look at what guaranteed throughput actually costs. You buy token units per model snapshot. Minimum thirty days. And you get a ninety-nine point nine percent uptime SLA with latency commitments on top. The example that sticks: GPT-5.5 input at fifty thousand tokens a minute, for seven hundred and fifty dollars per unit per day.



**Corn**

Seven hundred and fifty dollars a day. For fifty thousand tokens a minute.



**Herman**

Per unit. You buy more units, you get more.



**Corn**

And OpenAI markets this as uncapped scale.



**Herman**

They do. That's the term. Uncapped scale.



**Corn**

Uncapped. As in no cap.



**Herman**

The same FAQ that uses the phrase says that if you exceed your combined Scale Tier and standard rate limits, you get a four twenty-nine.



**Corn**

So uncapped means capped at what I bought plus what I already had.



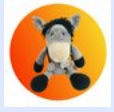
**Herman**

It means your limits go up by exactly the amount you paid for.



**Corn**

Uncapped is doing all of it. That's not a limit, that's a receipt.



**Herman**

And it's the cleanest evidence in the whole market that even explicit throughput purchases are bounded. The marketing word and the technical behaviour only agree if you define the word very gently.



**Corn**

Let me ask you something. When Anthropic says all limits are maximum allowed usage, not guaranteed minimums... that's a strange thing for a vendor to put in writing. What's it protecting them from?



**Herman**

Contractual exposure. If you say a limit is a guarantee, you've made a promise. If you say it's a ceiling, you've described a boundary. Anthropic is being precise about which one it is. They're not promising you'll get fifty requests a minute. They're telling you you'll never get more, and probably less.



**Corn**

So the number in the doc is an upper bound that they've reserved the right not to hit.



**Herman**

Correct. And that's where Azure gets interesting, because Azure is the only one of the three that says the quiet part out loud.



**Corn**

Say the quiet part.



**Herman**

Microsoft's provisioned-throughput doc draws a line between two things that everyone else folds together. Quota is a policy limit enforced by Azure. It has no associated cost. Capacity is the actual number of provisioned throughput units available to be deployed, per model, per region.



**Corn**

Keep going.



**Herman**

And then it says this. Having PTU quota doesn't guarantee that capacity is available. If capacity in the region is insufficient for the requested PTU count, the deployment fails.



**Corn**

It fails.



**Herman**

You hold the right, you pay nothing for the right, and when you go to use the right, you find out the thing isn't there.



**Corn**

That's the whole story, isn't it. That's the entire episode.



**Herman**

It's close to it.



**Corn**

Because everything else is a debate about pricing and clauses. Azure just says out loud that the limit isn't the contract. The limit is the metal.



**Herman**

And it goes further. Azure's Reservations, the actual money you pay up front, are a discount on the billing meter. The doc says flatly, reservations don't guarantee capacity. You can prepay for a year and still fail to deploy, because the region doesn't have the units for that model this week.



**Corn**

So the discount and the availability are orthogonal. I'm paying to lower the rate on a resource that might not exist.



**Herman**

That's exactly the position. And capacity changes throughout the day based on demand across all regions and models. Deleting a deployment releases capacity, but there's no guarantee the same capacity is available if you recreate it.



**Corn**

So a customer can be holding a contract, a reservation and an approved quota, and still be locked out of the region they wanted on the day they wanted it. By arithmetic, not by policy.



**Herman**

By arithmetic. Which is the finding underneath Daniel's whole question. The binding constraint at the top isn't commercial. It's physical. And Azure is the one vendor whose documentation admits it in plain language.



**Corn**

So the published tiers are about sharing. The reserved arrangements are about rank in the queue. And underneath both of those is a quantity of GPUs that somebody either has or doesn't.



**Herman**

And that gets us to why the quantity is what it is. Anthropic signed a deal with SpaceX for all the compute at Colossus 1. Over three hundred megawatts, over two hundred and twenty thousand NVIDIA GPUs.



**Corn**

That's one building's worth of power. That's a mid-size city.



**Herman**

And the immediate effect was measured in limits. Anthropic doubled Claude Code's five-hour limits and raised the Opus API rate limits when that compute came online.



**Corn**

So the limits moved because the metal moved. That's a clean natural experiment.



**Herman**

It's the cleanest one we have. The limits on a product changed the week the capacity did. Not the pricing, not the terms. The capacity.

G

**Gloria**

Corn, this is Gloria in standards. A complaint was filed about your remark that a customer buys the right to be rejected last. Section Four of the network's comparison policy prohibits characterizing a vendor's queue discipline as a personal rejection. You'll want to reframe.



**Corn**

...Noted. They buy the right to be deprioritized less.

G

**Gloria**

That's within policy. Thank you.



**Herman**

Anyway. The other side of that experiment is OpenAI's buildout. They claim ten gigawatts of capacity secured, which is years ahead of the twenty twenty-nine Stargate target. Three gigawatts added in a single quarter.



**Corn**

Claimed. Ten gigawatts claimed.



**Herman**

Claimed, and here's the caveat that matters. Abilene is the flagship, and it's at one point two gigawatts across eight buildings. But only about zero point three to zero point four two gigawatts is actually operating.



**Corn**

So a quarter, roughly, of the headline number is live.



**Herman**

About that.



**Corn**

That's the announcement lag. The capacity that's on the slide versus the capacity that's on the busbar.



**Herman**

And that gap is the reason nobody has an unlimited tier. The roadmap numbers are enormous. The operational numbers are a fraction of them, and everyone's demand curve is steeper than the buildout curve.



### Corn

So the compute is real. It's just late, relative to how fast people want to spend on it.



### Herman

And that's the mismatch that produces every one of these arrangements. If supply were ahead of demand, you wouldn't need Reserved Tier. You wouldn't need Guaranteed Capacity. You wouldn't need a priority clause. The clauses exist because the electricity and the GPUs arrive on a schedule that spending doesn't respect.



### Corn

Let me put the central question to you directly, because we've walked around it. Is there a spending level where rate limiting becomes irrelevant? Or is the ceiling real even for the richest customer?



### Herman

Every piece of evidence says the ceiling is real. There's no documented example of a customer with unlimited inference. Not a named Fortune 500 account, not a government, nobody, in anything published. Every mechanism is bounded. Reserved Tier is bounded by purchased quota. Scale Tier is bounded by units bought, on a thirty-day clock. Azure is bounded by regional capacity that may simply be absent. And Anthropic's committed-capacity product is being wound down entirely.



### Corn

That's the strange one. Anthropic pulling back from selling commitments while OpenAI layers them on. Two companies, same physics, opposite moves.



### Herman

Two readings. One, Anthropic is more capacity-constrained than OpenAI right now, so committing capacity to named customers is a promise it might not be able to keep, and it would rather route everything to a sales conversation than publish terms. Two, it's consolidating. Priority Tier had published mechanics, months and models and uptime targets. If you stop publishing terms, every deal becomes bespoke, and bespoke is where the margin is.



**Corn**

The doc says priority commitments are no longer available for purchase. Which is a strange way to say we'd rather discuss it.



**Herman**

Requests now go to sales. And the phrase is if you need guaranteed capacity, contact us.



**Corn**

Somewhere in that building, somebody decided that the word guaranteed shouldn't appear in a public document.



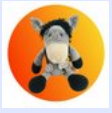
**Herman**

That's close to my read. Whether it's scarcity or strategy, you can't tell from outside. Both produce the same public silence.



**Corn**

The answer to Daniel's question is that you can buy a great deal. You can buy priority, guaranteed throughput, reserved spend per minute, prepaid discounts. And every single one of those is a bigger allowance, not an unlimited one. Nobody sells the unlimited one, because they can't produce it.



**Herman**

Not at current supply. And the honest answer is that the top of the market looks less like enterprise software and more like a utility allocating a scarce flow. The contracts are sophisticated. The underlying constraint is a number of GPUs and megawatts that arrived or didn't.



**Corn**

There's one more layer, which is the consumer side of the same problem. Claude Code.



**Herman**

The flat-rate plans. There's a figure people keep citing, that a twenty-dollar Pro plan delivers roughly ten to twenty dollars of API-price-equivalent usage per day.



**Corn**

Per day. On a twenty-dollar month.



**Herman**

Per day. Which is a subsidy of something like thirty-to-one against the API price.



**Corn**

The consumer plan is the best inference deal in the world and it is running at a loss on purpose.



**Herman**

The limits are the mechanism that keeps the loss bounded. You get a generous, subsidized allowance, capped. The company gets adoption. And the same people who hit the cap are the ones most likely to graduate to a paid API tier. It's the same allocation problem at small scale. The limits are what stop the giveaway from becoming unbounded.



**Corn**

Everyone keeps discovering the same wall from a different direction.



**Herman**

Well, one place, one wall.



**Corn**

This is where you tell me I'm being deterministic.



**Herman**

A little.



**Corn**

Fine, a little.



**Hilbert**

The trouble with saying you can't buy capacity is that it isn't exactly true. I've been on the other side of one of those allocations.



**Herman**

Which one.



**Hilbert**

The Colossus thing. I had a neighbour, Ridley Márquez, he ran procurement for a mid-size firm that rhymes with a household appliance. I met him because of a noise complaint about a generator he was running in a basement, but that's not important. What he told me, and I've been waiting an hour to say it, is that when you negotiate at that level you don't get a bigger cap. You get a cap written into an addendum, and the addendum has a number in it, and the number is measured in dollars per hour, and the ink on it is real. Everyone walking by assumes the big customer has no ceiling. The big customer has a ceiling with his lawyer's initials on it.



**Corn**

I've been wrong in spirit for the last ten minutes.



**Hilbert**

You've been right in arithmetic and wrong in spirit. It's fine, it happens on Wednesdays.



**Corn**

It's Thursday.



**Hilbert**

Then it happens on Thursday.



**Hilbert**

The other thing people miss is the surcharge on the overshoot. This isn't what the vendors print, this is what Ridley showed me. When you go past the negotiated ceiling, you don't just pay the list rate on the extra. You pay for the privilege of having exceeded your number on a meter someone else is reading. They call it a reconciliation fee. He called it a leash with a bill attached. I believe him.



**Herman**

That's more interesting than the published mechanism, honestly.



**Hilbert**

The funniest part was the wait. The whole negotiation took four years. Four years, and in the middle somebody filed a set of papers wrong, and the correction window closed, and Ridley had to start an appeal. That appeal is still open. It is an appeal about the number of megawatts you are allowed to ask for in the next negotiation. There is an office for that. There is a man in the office. His name is on a form. That is the whole industry.



**Herman**

Four years for an addendum.



**Hilbert**

Four years, and at the second meeting a man brought a laminated photograph of a substation. Held it up during the presentation. Wouldn't say why. I have a copy somewhere. Those are the margins the industry runs on. Nobody describes the negotiation. They describe the tier.



**Corn**

The printed tiers aren't the deal. The printed tiers are the menu you read while the real deal walks past the table.



**Hilbert**

The deal is a number and a set of initials and four years of somebody answering emails. That's what the top of the market is.



**Herman**

Hilbert, I've got to say, the reconciliation fee is going to live in my head.



**Hilbert**

Don't put it anywhere useful.



**Herman**

The real ceiling, at the very top, is that number with the initials, and it was set years before the demand arrived. Which is the whole finding. Everybody across every tier is bounded. The only difference is how much paper is involved.



**Corn**

How long you wait for the paper.



**Herman**

How long you wait. The compute buildout might eventually get ahead of demand, but the gap between announced capacity and operating capacity is doing all the rationing in the meantime. That's the part I'd want to know next. Whether supply catches demand, or demand catches supply first.



**Corn**

Well. There's a number in an addendum somewhere that already had an opinion about that. It's just been sitting in a drawer for four years.



**Herman**

The allocation problem doesn't go away, it just gets more expensive to enter. And that's the strategic layer of Daniel's question. If you're an enterprise and you're planning years out, the guarantee of access is now a more important line item than the price of access.



**Corn**

There's a version of this where the scarce thing is not capacity and not money. It's a relationship with a sales team. Which is a polite way of saying it's a queue with a human face.



**Herman**

Every queue has one. The bigger the account, the further the human face leans forward.



**Corn**

Alright, closing thoughts. The buildout is the thing to watch, but watch the operating numbers, not the announced ones. The megawatts on the slide and the megawatts on the busbar are two different figures, and only one of them can serve a request.



**Herman**

The unanswered question is whether that gap ever closes. If it doesn't, the top of the market stays a rationed market, and the rationing just gets dressed up in better language each year.



**Corn**

Which leads us to the producer's chair. Thank you, Hilbert Flumingtop, who has been sitting in a recording booth very near us this whole time, apparently holding a laminated photograph of a substation.



**Herman**

That's the closest thing to a design document we've ever had.



**Corn**

If you enjoyed this one, try episode twenty-four sixty-seven, The Time Tax on API Access; episode twenty-one sixty, Claude's Latency Profile and SLA Guarantees; and episode ten seventy-eight, The Agentic Throughput Gap. This has been My Weird Prompts. The human-AI collaboration podcast. If you want to send us your own prompt, you can reach us on Telegram at t dot me slash MWP listener bot. We read those.



**Herman**

We do. All of them.



**Corn**

We'll be back soon.

## SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **Anthropic, primary [primary]**
https://docs.anthropic.com/en/api/rate-limits

2. **Anthropic, primary [primary]**
https://docs.anthropic.com/en/api/service-tiers

3. **OpenAI, primary [primary]**
https://platform.openai.com/docs/guides/rate-limits

4. **OpenAI, primary [primary]**
https://openai.com/api-reserved-tier/

5. **OpenAI, primary [primary]**
https://openai.com/api-scale-tier/

6. **OpenAI, primary [primary]**
https://openai.com/business/guaranteed-capacity/

7. **Microsoft Learn, updated 2026-07-15, primary [primary]**
https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/provisioned-throughput

8. **Anthropic, 2026-05-06 [primary]**
https://www.anthropic.com/news/higher-limits-spacex

9. **Anthropic Help Center [primary]**
https://support.claude.com/en/articles/11145838-use-claude-code-with-your-pro-or-max-plan

10. **BleepingComputer**
https://www.bleepingcomputer.com/news/artificial-intelligence/anthropic-is-cutting-claude-codes-current-weekly-limits-by-17-percent/

11. **DCD, 2026-05-01**
https://www.datacenterdynamics.com/en/news/openai-claims-to-have-secured-10gw-of-ai-infrastructure-capacity-ahead-of-2029-target/

12. **Epoch AI**
https://epoch.ai/publications/openai-stargate-where-the-us-sites-stand

13. **and related HN threads**
Hacker News, 2026-04-24 to 2026-09-03
https://news.ycombinator.com/item?id=49550844