Episode Book

EPISODE #6003

What Enterprises Actually Pay for AI Inference

PUBLISHED	15 October 2026 at 18:39 UTC
DURATION	25:19
LISTEN	myweirdprompts.com/episode/enterprise-ai-inference-spend-79972/
HOSTS	Corn, Herman, Raj, 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

Enterprise spending on frontier model APIs hit $12.5 billion — yet no lab publishes who actually pays what. This episode digs into the brackets, from self-service startups spending hundreds a year to multinationals with nine-figure commitments, and explains why the scariest numbers in circulation are compute purchases, not customer bills. Then it turns to what generates the bill in the first place: coding agents that retry until they work, seat-based copilots, and the surprising fact that only 16% of enterprise deployments are true agents. If you've ever wondered why AI bills climb while token prices fall, this is the episode.

TRANSCRIPT



Corn

Twelve and a half billion dollars. That's what every company on earth spent on frontier model inference last year. And nobody can tell you who spent it.



Herman

The single most requested number in enterprise AI, and it arrives with a hole in the middle of it.



Corn

Daniel's been reading about the infrastructure side of this, the serving question, and now he wants the money side. His words: put some actual dollar figures on enterprise consumption of proprietary models. GPT, Claude, the frontier labs.



Herman

He wants the brackets. Startups at one end, Microsoft at the other.



Corn

He wants to know what a typical organization actually pays, how far that spreads, and which public data — if any — tells us how many organizations sit in each category.



Herman

And then the interesting half.



Corn

What generates the bill in the first place. Employee use, customer-facing products, coding agents, automated workflows. And one thing he's firm about: keep API inference separate from software subscriptions, and separate again from training.



Herman

Three buckets, and the headline numbers mix all three.



Corn

Which is why the headline numbers are useless.



Herman

The whole product line spans nine orders of magnitude. You can open an account this afternoon and pay four cents for a million tokens. You can also be a company with a nine-figure annual commitment to the same product family. Same pricing page.



Corn

Same pricing page, different planet.



Herman

So let's set the denominator first, because everything else hangs off it. Menlo Ventures does the annual enterprise survey, and their 2025 report puts total enterprise generative AI spend at thirty-seven billion dollars. That's up from eleven and a half billion in 2024. A little over three times.



Corn

Thirty-seven billion total.



Herman

Of that thirty-seven, eighteen went to infrastructure. And inside the eighteen, twelve and a half billion was foundation model APIs. That twelve and a half is the number. That's the closest thing anyone has to a hard figure for aggregate enterprise API inference spend.



Corn

The actual question Daniel asked, in other words. The market for the thing.



Herman

The other nineteen billion went to applications. Copilots, coding tools, vertical AI. That's largely seat-based subscription spend, and it has to stay separate or you double-count. And training infrastructure is a further four billion, also separate.



Corn

So Daniel's boundary lines are exactly right. Twelve and a half is the API line, nineteen is the software line, four is the build-it-yourself line.



Herman

Three different businesses that get reported as one number whenever a lab announces revenue.



Corn

And the twelve and a half is where the public data starts getting thin. Almost immediately.



Herman

Immediately. Neither lab publishes a histogram of customers by annual API spend. Not OpenAI, not Anthropic, not Google. The market size is published. The share is published. The distribution underneath it is not.



Corn

So everything after this point is inference from disclosed floors and reported contracts. Not a published distribution.



Herman

There is exactly one disclosed cut. Anthropic said in April that more than a thousand customers spend over a million dollars a year. And that number had doubled from five hundred in about two months.



Corn

A thousand customers above a million. That's a floor, not a ladder.



Herman

It's a floor. It tells you the top of the pyramid is real and it's growing fast. It tells you nothing about the shape below it. And Anthropic also reports three hundred thousand plus business customers as of September 2025.



Corn

Hold on. Three hundred thousand customers, and a thousand of them are above a million a year.



Herman

Do the arithmetic on the total. Twelve and a half billion across three hundred thousand business customers is a mean of roughly forty thousand dollars a year. But nobody publishes a median. And the mean is doing something dishonest there.



Corn

Because one customer at fifty million drags up a hundred small ones.



Herman

If a thousand customers are above a million, and some of those are well above, you can account for a very large share of the twelve and a half billion with a few hundred accounts. Which means the median is far below the mean. My guess is the median enterprise API spend is a few thousand dollars a year.



Corn

A few thousand, and it's the number that describes most customers.



Herman

Most customers by count. By dollars, the distribution is upside down. It's a barbell with a very thin handle.



Corn

So give me the brackets, then. Clearly labeled as estimates.



Herman

Estimates, all of them. At the bottom, self-service startups and small teams. Hundreds of dollars a year to low thousands. That's consistent with OpenAI's published tiers, where you can run a real workload for the price of a lunch. Then medium businesses and scale-ups, where you're past the free tier and into committed spend. Tens of thousands to low hundreds of thousands a year.



Corn

That's where most of the three hundred thousand sit, I'd guess.



Herman

Probably. Then smaller Fortune 500 companies and large enterprises. Low millions. They've got a procurement process, they've negotiated a rate card, they've got a committed spend contract. Then the enormous multinationals, which is tens of millions to hundreds of millions.



Corn

And the top of that bracket is Microsoft, per Daniel's prompt.



Herman

Microsoft is the interesting one, because the biggest number associated with Microsoft and AI isn't an inference bill at all. Executive testimony in May put Microsoft's cumulative spend on the OpenAI partnership at over a hundred billion dollars.



Corn

On the partnership.



Herman

Building capacity, funding the relationship, the whole arrangement. That is not a bill for tokens. It's the cost of owning the supplier.



Corn

That's a completely different line item wearing the same headline.



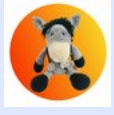
Herman

And it's the mistake the coverage makes constantly. Anthropic has a five hundred eighteen billion dollar decade-long compute commitment. Roughly eighty percent non-cancelable. Two hundred billion to Google Cloud over five years. Fifteen billion a year to SpaceX through May 2029. Those are compute purchases. Anthropic is buying capacity, not receiving an invoice for inference.



Corn

So the scariest numbers in circulation are Anthropic's shopping list.



Herman

Its shopping list. And there's a second confusion stacked on top. What the labs pay each other and what customers pay the labs are different flows entirely. OpenAI paid Microsoft seventeen point two billion dollars in 2025. Ten and a half of that was R&D and training, six billion was cost of revenue, which is inference, and about half a billion was sales and marketing.



Corn

OpenAI paying Microsoft.



Herman

Meanwhile Microsoft paid OpenAI three hundred and three million dollars. Fifty-six to one in OpenAI's direction. Neither of those is a customer's bill. That's two companies settling up for compute and for a partnership.



Corn

So we've got a market of twelve and a half billion, one disclosed floor at a thousand customers above a million, and no verified figure — anywhere — for what a single organization actually pays in a year for inference.



Herman

Not one. The closest thing is OpenAI's own cost of revenue, which was seven and a half billion in 2025. That's what OpenAI spent serving everything, not what any customer spent.



Corn

Every number that sounds like an answer is a lab's cost, not a customer's bill.



Herman

Which is the honest summary of part one.



Corn

So the brackets are guesses, the distribution is undocumented, and the biggest figures people quote are compute purchases. That's the state of the public record.



Herman

The state of the public record is a denominator and a floor. Everything between them is estimation.



Corn

Then the more useful question. What actually generates the bill?



Herman

Coding. Coding is the first real killer use case in the enterprise, and it's not close. Enterprise coding spend was four billion dollars in 2025. The year before it was five hundred and fifty million.



Corn

Eight times in a year.



Herman

And that four billion is fifty-five percent of all departmental AI spend. More than half of what enterprises buy, at the departmental level, is coding tools.



Corn

Why coding? Why not customer service, why not legal?



Herman

Because of what a coding agent actually does per task. It makes many model calls for a single job. It reads files, it reasons, it edits, it runs the tests, it fails, it retries. Often with large context windows. Compare that to a chat copilot, which is one call and one answer.



Corn

So the same seat generates a very different bill depending on what it's doing.



Herman

An order of magnitude different, probably more. A copilot seat is a fairly predictable monthly cost. A coding agent is metered, and its meter runs every time it has a thought.



Corn

Including the wrong ones.



Herman

Especially the wrong ones. A retry loop is a billing event. And that's why Anthropic's enterprise lead is a coding story. It holds fifty-four percent of coding against OpenAI's twenty-one. That's the whole reason it took forty percent of enterprise API spend in 2025, up from twenty-four the year before and twelve the year before that.



Corn

And OpenAI went from fifty percent in 2023 down to twenty-seven.



Herman

Google's at twenty-one. Those three are eighty-eight percent of enterprise API spend between them.



Corn

Cheap to be the fourth one.



Herman

Now the horizontal copilots, the general assistants. That's seven point two billion, and it's eighty-six percent of the horizontal AI category. Big, real, but seat-based. A predictable per-head number, and it scales with headcount, not with ambition.



Corn

That's a budget line a CFO can live with.



Herman

The metered one is not. And here's the part that surprised me. Only sixteen percent of enterprise deployments are true agents. Sixteen percent. The other eighty-four percent are fixed-sequence workflows wrapped around a single model call.



Corn

So the agentic fleet everyone's writing about is a rounding error on the bill.



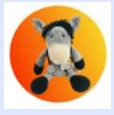
Herman

Today's bills are seat-based copilots and coding assistants. The autonomous multi-step thing that runs unattended — that's a small slice. It's growing, but it is not what's generating the twelve and a half billion.



Corn

Which Daniel asked about directly. Automated workflows.



Herman

Automated workflows are mostly fixed-sequence. Do this, then this, then stop. That's cheap per run. The expensive shape is the one that decides for itself how many times to try.



Corn

So where does that leave employee use versus customer-facing products?



Herman

Employee-facing tools are the bulk. Copilot seats and coding assistants, pointed at your own staff. Customer-facing AI products are growing and they're real, but they're a smaller share of the twelve and a half. And the reason is that a customer-facing product has a margin problem — every call a customer makes costs you money directly, so you're motivated to keep the model small and the calls few.



Corn

The internal tool has no such discipline, because the value is harder to measure.



Herman

There's no revenue line attached to a coding agent. There's just a bill.



Corn

So now the question underneath Daniel's question. Why do the bills climb when the price per token keeps falling?



Herman

This is the structural part, and it's the best thing in the whole picture. OpenAI's gross margin fell from about forty percent to thirty-three percent, while revenue tripled.



Corn

Revenue triples, margin falls.



Herman

Because the mix shifted toward reasoning models, and a reasoning model burns far more compute per query. It generates a chain of thought before it answers. Some of these models will spend thousands of tokens thinking about a question the user asked in twenty.



Corn

And the customer sees one answer.



Herman

And pays for the thinking. So you get a compound effect. Price per token falls, but tokens per task rises faster. Coding agents make many calls, reasoning models burn more per call, context windows keep growing because everybody wants the model to see the whole file. The meter you're reading got cheaper. The number of meters running went up.



Corn

And that's why the twelve and a half billion climbs even though the pricing page looks generous.



Herman

Every efficiency gain gets reinvested in a harder task. That's the pattern.



Corn

What about the billing path? Who actually invoices the customer?



Herman

Forty-seven percent of Anthropic's 2025 revenue was billed by Amazon and Google. Up from thirty-two percent in 2024. So when an enterprise says it's spending two million a year on Claude, close to half the time that's an AWS or a Google Cloud line item.



Corn

On an Amazon invoice.



Herman

On a cloud invoice, and the cloud owns the relationship. It knows the customer, it holds the contract, and it will happily sell you Gemini or a Llama endpoint off the same shelf.



Corn

So the labs are partly renting their customers.



Herman

And the customer gets a single cloud bill with the model line buried in it. Which is another reason no one can produce a clean figure for what a company spends on inference. It's inside a cloud bill that nobody breaks out.



Corn

The last piece. How safe is that revenue?



Herman

Anthropic's draft prospectus says two customers each accounted for twelve percent of 2025 revenue. And it notes that many large customers lack long-term commitments.



Corn

So the biggest bills are re-won every month.



Herman

Consumption pricing means the customer decides each month how much to consume. There's no three-year lock-in on the usage side. You can have a nine-figure account and lose it in a quarter if the team moves to a different model.



Corn

That's a strange business. Enormous revenue, enormous concentration, renewed monthly.



Herman

The Decoding Discontinuity teardown put it well. Their revenue is rent. Mostly metered and re-won month by month. Their costs are leases.



Corn

Rent in, leases out.



Herman

And the leases are the five hundred eighteen billion. That's the asymmetry. The upside is meters running in a customer's cloud account. The downside is a decade of non-cancelable commitments.



Corn

There's a detail in that teardown I keep coming back to. The revenue is concentrated in the customers who can leave fastest. The two accounts at twelve percent each — those are the ones with the leverage to renegotiate every quarter. So the stability of the whole business rests on the least stable part of it.



Herman

Which is why the labs push so hard on multi-year commits at the top. The smaller accounts are sticky by inertia. The large ones are sticky only by contract, and the contracts are short.



Corn

So the concentration risk and the churn risk are the same risk.



Herman

They always were. It just shows up in two different lines of the filing.



Corn

Raj, you've been sitting there the whole time. What have you got?

R

Raj

Gentlemen, this is not a market. This is a screenplay. Three hours, two intervals, one villain — the meter. I am seeing Herman as the numbers man, very strong, very strong, and Corn as the conscience, the one who asks the question nobody wants to answer. We open on the pricing page. Four cents a million tokens. Beautiful. Then smash cut to the boardroom, and the bill is eleven feet long. Interval twist — the cloud was the customer all along.



Herman

Raj, when you say the cloud is the customer, do you mean the billing relationship, or ownership of the account?

R

Raj

Both, neither, it doesn't matter, we fix it in the edit. And here is the part I love. We put a song in the retry loop. The agent tries again, and again, and each time the orchestra gets bigger. By the ninth retry the whole Mumbai Philharmonic is on its feet.



Corn

So the bill is the score.

R

Raj

My friend. And the last shot, on the CFO's face, as the monthly invoice arrives. No words. Just the invoice, and one tear, and the music swells. I have the option papers in my bag. Excuse me, another booking.



Herman

The orchestra gets bigger every time it fails. That's roughly the coding-agent economics, honestly.



Corn

So the picture is this. A thousand customers above a million, three hundred thousand customers total, and a median that's probably a few thousand dollars a year.



Herman

And the twelve and a half billion is carried by a small number of enormous bills. The tail isn't the tail. The tail is the market.



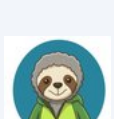
Corn

One more thing worth having for Daniel's list. The two labs' headline numbers aren't comparable.



Herman

The fifty billion versus seventy billion gap that was in the news last week. That's a basis difference, not a restatement. OpenAI counts what it keeps, net. Anthropic grosses up the sales billed through the clouds.



Corn

So one company's number includes money that goes straight to Amazon.



Herman

And the other's excludes money that goes to Microsoft. You cannot put those two figures side by side and declare a winner. It's an accounting convention wearing a scoreboard.



Corn

Every "who's winning enterprise AI" chart is built on that.



Herman

Every one I've seen.



Corn

Herman, you told me you'd read something overnight that you weren't going to bring up.



Herman

I did, and I'm going to bring it up. Sacra's estimate has OpenAI's inference costs going from eight point four billion this year to fourteen point one billion next year.



Corn

So the cost line grows faster than the customer count.



Herman

Substantially. And that's before the agentic share moves off sixteen percent.



Corn

Which it will.



Hilbert

How much does a token cost when nobody's watching it?



Herman

That's the whole problem in one line.



Corn

Hilbert, I think you've solved the episode.



Hilbert

I didn't ask it to solve anything. I asked it because I've been on the phone with my nephew since you started talking, and he's got a number for me, and the number is not what he said.



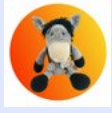
Corn

He works in procurement, then.



Hilbert

He works in procurement at a hospital group. Twelve hospitals, all regional, all paper files. Four years ago they set up their own department for the AI line items. Two people. He's one of them. His job is to go through the bills and find out what's actually on them.



Herman

What did he find?



Hilbert

Their annual invoice from the model provider was sixty-eight thousand dollars and change. Sixty-eight thousand four hundred and something. He read it twice because he thought the decimal was in the wrong place. It was his first month on the job. And the number that mattered wasn't the sixty-eight thousand. It was what sat next to it on the same statement, from the same cloud, for storing the files the model reads. That was two hundred and eleven thousand.



Corn

The storage cost more than the inference.



Hilbert

The file storage was two hundred and eleven thousand and the model was sixty-eight. So he goes to his director and says, we're paying three times as much to keep the files as we are to think about them. And the director says, well, that's the storage line, that's a different budget.



Herman

Different budget. Same invoice.



Hilbert

Different budget. His department can only see the sixty-eight. The storage sits under IT infrastructure in a code he doesn't have access to. Two years he's been asking for the code. Every quarter the request gets closed as resolved with no note. He now knows the number by heart and he cannot get a copy of the line. His manager told him to stop raising it in meetings because it makes the AI programme look bad.



Corn

The storage is doing the honest accounting his department won't.



Hilbert

He asked me to guess the total for all twelve hospitals and I gave him a figure and he laughed and said it was eleven percent high, which he knew without looking it up.



Herman

Which side of it is inference, from his point of view?



Hilbert

None of it. In his filing it's one line item called "cloud services" and it's been the same code since before the AI programme existed. That's why he can't find the model spend and the storage spend separately. They were never separated. He's reconstructing it from purchase orders.



Corn

The four years, that's the whole time the department's existed.



Hilbert

Four years, two people, and a code he can't get.



Herman

The sixty-eight thousand number is what he pays for the model...



Hilbert

That's one hospital group's number and it's already buried in a code nobody can see.



Corn

The only clean figure any of us has heard this whole episode came from a man reconstructing it from purchase orders.



Herman

It's smaller than the storage sitting next to it. Which changes the shape of Daniel's question a little. If the storage line routinely outruns the model line, then the model spend isn't where the argument is. The argument is that it's invisible.



Corn

Hilbert, the director — did he ever look at the sixty-eight thousand?



Hilbert

He called it encouraging. Said it showed the programme was disciplined. He's not wrong. It's the storage he can't see.



Herman

Right. Because the inference metric everybody's building charts about is the smallest number on that statement.



Corn

The smallest number, and the only one anyone asked about.



Hilbert

He sent me the invoice as a photo. Four pages. He's keeping it in a drawer.



Corn

A drawer somewhere in a hospital group, and that's the state of enterprise inference accounting.



Herman

That's the most honest figure in the episode. Sixty-eight thousand, and it's smaller than the thing next to it that nobody counts as AI.



Corn

Which puts the twelve and a half billion in a new light. The aggregate is real, the bills are real, but a huge amount of enterprise model spend is sitting in a code called cloud services, next to storage, and the only person who's looked at it closely had to fight for the access.



Herman

When the next earnings call cites enterprise momentum, that number includes a lot of line items that were never labelled.



Corn

Last thing before we close, because Daniel asked for the most credible picture, and this is it. There's a gap between the two labs' headline revenue figures that isn't a gap.



Herman

Fifty against seventy. One of those numbers is net of what the cloud takes. The other grosses it up. Put them side by side and it looks like a race. It's two different rulers.



Corn

The "who's winning" question is answerable and unanswerable at the same time.



Herman

Answerable on share, where they're measured the same way. Unanswerable on the headline, where they're not. And the headline is the number everyone quotes.



Corn

Which is why Daniel's instinct to separate the buckets was right. Subscriptions, training, inference. Separate them and the picture gets smaller and truer. Add them together and you get a number you can't defend.



Herman

Small and true beats big and blurry.



Corn

Where does the bracket structure go from here?



Herman

Up, if the per-query compute keeps rising and the agentic share moves off sixteen percent. The bottom of the distribution probably stays where it is. The top gets further away. Which is the shape of every enterprise software category, honestly.



Corn

The bottom stays a few thousand dollars a year and the top stays a small number of enormous accounts. The brackets don't move. The gap inside them widens.



Herman

The median is a few thousand dollars and mostly unpublished, which is the least satisfying answer we've ever given anyone.



Corn

It's the honest one. The most credible picture is that the twelve and a half billion is carried by a handful of very large bills, that the median enterprise spends almost nothing by comparison, and that the mega-contracts in the headlines are compute purchases, not inference revenue.



Herman

Which is what Daniel was really asking. What does an organization actually pay, and the answer is it depends on which organization, by four or five orders of magnitude, and nobody publishes the ladder.



Corn

The storage next to the model probably costs more.



Herman

In at least one hospital group it does.



Corn

Thanks to Hilbert Flumingtop, our producer, for his nephew and his drawer.



Herman

For the sixty-eight thousand.



Corn

For more along these lines, there's episode forty-eight, Renting vs. Building; episode three, Safetensors or something else; and episode ten, How ASR Went From Frustration To ... Whisper Magic. This has been My Weird Prompts, the human-AI collaboration podcast. If you want to put a figure on your own inference bill, or you've spotted one we can't, send us your own prompt on Telegram at t dot me slash MWP listener bot. We read everything.



Herman

We do. And if you enjoyed this one, a review helps more than you'd think.



Corn

See you soon.

SOURCES

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

1. **OpenAI, current [primary]**
   https://openai.com/api/pricing/

2. **Menlo Ventures, 2025-12-09 [primary market research]**
   https://menlovc.com/perspective/2025-the-state-of-generative-ai-in-the-enterprise/

3. **Decoding Discontinuity, 2026-10-13**
   https://www.decodingdiscontinuity.com/p/openai-anthropic-revenue

4. **Ed Zitron, 2026-06-15 [primary audited figures]**
   https://www.wheresyoured.at/exclusive-openai-financials/

5. **TechTimes, 2026-06-16**
   https://www.techtimes.com/articles/318496/20260616/openai-lost-385-billion-2025-audited-financials-expose-17b-azure-dependency.htm

6. **CNBC, 2026-08-17**
   https://www.cnbc.com/2026/08/17/anthropic-says-annualized-revenue-climbed-to-65-billion-in-july.html

7. **Axios, 2026-10-08**
   https://www.axios.com/2026/10/08/openai-50-billion-arr-anthropic-revenue-accounting

8. **Reuters via Investing.com, 2026-10-08**
   https://www.investing.com/news/stock-market-news/openais-annualized-revenue-20-billion-less-than-previously-signaled-ft-reports-4939854

9. **InvestmentNews, 2026-09-29**
   https://www.investmentnews.com/equities/anthropics-landmark-ipo-filing-shows-12-fold-revenue-jump-518b-compute-bill/268391

10. **Axios, 2026-05-20**
    https://www.axios.com/2026/05/20/anthropic-spacex-compute

11. **Bloomberg, 2026-08-05**
    https://www.bloomberg.com/news/articles/2026-08-05/microsoft-s-ai-sales-mostly-come-from-openai-disclosures-show

12. **AI Tech Model, 2026-10-05**
https://aitechmodel.com/bull-case-vs-bear-case-for-anthropic-enterprise-revenue-mix/

13. **The Backfield River (OpenAI/Microsoft ledger detail)**
https://backfield.net/river/card/6625