Episode Book

EPISODE #6217

Fine-Tuning a Trillion-Parameter Model at Home

PUBLISHED	22 October 2026 at 23:34 UTC
DURATION	21:24
LISTEN	myweirdprompts.com/episode/trillion-parameter-fine-tune-cost-60285/
HOSTS	Corn, Herman, Dorothy, Hilbert
SCRIPT MODEL	anthropic/claude-sonnet-5.5
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.20
BOOK GENERATED	23 October 2026 at 07:44 UTC

SYNOPSIS

Daniel fine-tuned a speech model on his own mixed Hebrew-and-English recordings for the price of a coffee. So what happens if you point that same move at an Opus 5.5-class frontier model — raw, full precision, no practical reason at all? This episode walks through the arithmetic: the sixteen bytes per parameter that Adam demands, why a one-trillion-parameter model needs roughly 200 H100s just to hold its training state, and why mixture-of-experts saves compute but not memory. Along the way: what Modal's eight-GPU ceiling actually buys you, why a ten-billion-token fine-tune costs about $166,000 in pure compute, and what 1,024 GPUs and 1.8 megawatts of power look like as a physical room.

TRANSCRIPT



Corn

What gets me is how undramatic it was. You press go, you wait roughly as long as lunch, and a speech model that didn't really cope with Hebrew and English in the same sentence now sort of does. Daniel did that today with Parakeet, on his own mixed-language recordings, and it cost him a few dollars.



Herman

Twenty to thirty minutes of GPU time. The first run didn't fully work, which is the normal condition of a first run, but the weights moved. That's the part I can't get over either. Somebody else's model, and now it's slightly his.



Corn

And he'd like it said on air that he is a huge fan of Modal.



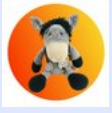
Herman

Deservedly. You write a Python function, you say which GPU you want, and you pay by the second for the seconds you used. An H one hundred is about four dollars an hour at their list price, so a half-hour run is a coffee. Ten years ago that was a grant application and a month of begging the department for cluster time.



Corn

Ten years ago it was a man called Gary who controlled the cluster, and you brought Gary biscuits.



Herman

Gary is still out there, somewhere. Anyway.



Corn

So here's Daniel's question. Take that same move, shifting the weights, and point it at the biggest thing there is. Not Parakeet, not a quantized model with a small adapter clipped on the side, but a full fine-tune of an Opus five point five class model. Raw. Full precision. No practical reason at all. He wants to feel the gulf, in GPUs and machines and hours and money, and in what the room sounds like.



Herman

First wrinkle, and it's a fun one. Anthropic doesn't say how big Opus five point five is, and nobody outside Anthropic can fine-tune it, so we get to choose a size.



Corn

A licence to be wrong with confidence.



Herman

With anchors, though. The open models at the top end tell us what a frontier model of this era looks like. Kimi K two is around one trillion parameters total, with thirty-two billion active per token, because it's a mixture of experts. DeepSeek V four Pro is one point six trillion total, forty-nine billion active. The biggest open one I've seen reported is Kimi K three, around two point eight trillion with a hundred and four billion active, though I'd treat that number as blog-level until I see a model card. So for the show: one trillion dense, two trillion dense, and a one-to-three trillion mixture of experts.



Corn

Dense meaning every parameter does work on every token.



Herman

Right, and the difference is going to matter a lot, because it splits the bill into two bills. One for memory, one for compute. Let's do memory, because it's the one people forget, and it's where the gulf opens.



Corn

Go on. I'll try not to fall asleep at the arithmetic.



Herman

Parakeet first. About six hundred million parameters. When you train a model with the standard Adam optimizer in mixed precision, each parameter costs you sixteen bytes while you're training. Two bytes for the weight in bfloat sixteen. Two for its gradient. Then Adam wants a full-precision master copy of the weight, four bytes, and two running averages, four bytes each. Two, two, four, four, four. Sixteen.



Corn

So I'm paying sixteen bytes to hold something that ends up as two.



Herman

Yes, which is why a quantized fine-tune is such a bargain, but we're doing the raw one. Parakeet comes out around ten gigabytes of training state. That's one modest GPU with room to spare, plus the activations, which depend on batch size and are on top.



Corn

Ten gigabytes. I have photos bigger than Daniel's entire training run.



Herman

Now the same sixteen bytes, on one trillion parameters. Sixteen terabytes. Before a single token goes through it.



Corn

Hm. And an H one hundred holds eighty gigabytes.



Herman

So two hundred cards, just to hold the state. Not to compute. To hold. And real runs need maybe a third to a half more for activations and communication buffers, so call it three hundred H one hundreds for the one-trillion model. Two trillion is thirty-two terabytes, four hundred cards raw, five or six hundred in practice.



Corn

Now the obvious cheat. The quantized route, the adapter clipped on the side. What does that do to the sixteen terabytes?



Herman

It's a different universe. Freeze the base model at four bits, so half a byte per parameter, and train only a thin adapter on top. One trillion parameters is then roughly half a terabyte of frozen weights. By my arithmetic that's single digits of big cards, and one box if the box is a B three hundred. Sixteen terabytes against half a terabyte. Same model, thirty-odd times less to hold.



Corn

So the gulf isn't between small and big models. It's between raw and clever.



Herman

A lot of it, yes. But hold the thought, because we're doing the raw one.



Corn

Wait. Mixture of experts is the clever one, though. Thirty-two billion active. Doesn't that make it cheap?



Herman

That's the misconception, and I'd have guessed the same a few years ago. Memory scales with total parameters, not active ones. Every expert has to be sitting in memory, with its own gradients and its own optimizer state, because any of them might get picked on the



Herman (continued)

next batch. The mixture of experts saves you compute, not memory.



Corn

So the one trillion mixture of experts still needs two hundred cards of space.



Herman

Same sixteen terabytes. Newer cards help, to be fair. A B two hundred holds a hundred and ninety-two gigabytes, so the raw count drops to about eighty-three for one trillion. A B three hundred, two hundred and eighty-eight, so fifty-six. But you're still talking about seven to ten machines minimum, and no one runs at the minimum.



Corn

So how far does the platform Daniel loves actually get you?



Herman

I was hoping you'd ask, because the answer is a clean illustration. Modal gives you up to eight GPUs in one container, which is one physical machine. Eight B two hundreds is about one and a half terabytes of memory, and at sixteen bytes a parameter that tops out around a ninety-five billion parameter full fine-tune. Eight B three hundreds, maybe a hundred and forty-five billion. Multi-node is in private beta, as far as I can tell.



Corn

So the fan's limit is about a tenth of the trillion-parameter model.



Herman

About a tenth, on the biggest machine they rent. A full fine-tune of Llama four hundred and five billion, six and a half terabytes of state, doesn't fit on one.



Corn

And I'm fine with that. Nobody made that platform for this.



Herman

Nobody did. Now, you split the state across many GPUs, so each holds a slice. That's what the sharding schemes do, and then you add tensor parallelism inside each machine, pipeline parallelism across machines, expert parallelism for the mixture. For the biggest Llama pretraining run, the team combined all of those across sixteen thousand cards.



Corn

Herman, I'm going to guess. You brought the whole paper.



Herman

I brought a respectable portion of the paper.



Corn

Then compute. What does the training cost, with the cards in hand?



Herman

A good rule of thumb is six times the active parameters times the tokens, in floating-point operations. Assume forty percent of the card's peak, which is roughly what the Llama three team reported. On an H one hundred at Modal's list price, a ten-billion-token fine-tune of a one-trillion dense model is about forty-two thousand GPU-hours. Call it a hundred and sixty-six thousand dollars. Two trillion dense, double: about three hundred and thirty thousand.



Corn

And Daniel's run, in the same units?



Herman

My estimate, since I don't know his data size: a hundred hours or so of audio, a few passes over it, comes to the order of ten to the seventeenth operations. That's minutes at the card's peak. So most of his half hour was probably loading, shuffling and saving, not arithmetic. A ten-billion-token run on the trillion-parameter model is about six times ten to the twenty-second. Five or six orders of magnitude apart.



Corn

Ten billion tokens is big for a fine-tune, though.



Herman

It is. Most fine-tunes are a tiny fraction of that. At one billion tokens, divide by ten: sixteen or seventeen thousand dollars for the one-trillion dense. At a hundred billion, which is a serious continued-training job, multiply by ten: one point six six million.



Corn

So Daniel's few dollars, and then a few hundred thousand, and then a few million. Depending on how much you feed it.



Herman

And the mixture of experts is where it gets funny. One trillion with thirty-two billion active, ten billion tokens: roughly thirteen hundred GPU-hours of pure compute. Five thousand dollars. A thirtieth of the dense bill.



Corn

Five thousand dollars for the weights of a one-trillion-parameter model. That's cheap!



Herman

Careful. That's the compute alone, at a perfect forty percent. You still need two hundred cards just for the memory, and tokens have to be routed between experts across machines on every layer, so utilization is worse. Commonly seen figures are twenty to



Herman (continued)

thirty percent for mixture-of-experts at scale, though I don't have a clean source for that. Your five thousand becomes eight or ten, and you've rented a small data centre to spend it.



Corn

Small bill, enormous room.



Herman

Memory-bound and communication-bound. Your five thousand dollars buys the arithmetic and none of the logistics.



Corn

And wall-clock? If I have a thousand of them.



Herman

A thousand and twenty-four H one hundreds, one-trillion dense, ten billion tokens: about forty-one hours. Two trillion: eighty-two. The hundred-billion-token version, about seventeen days.



Corn

Seventeen days. I've had naps that long.



Herman

For scale, the four-hundred-and-five-billion Llama pretraining used sixteen thousand cards for about fifty-four days. That's something like twenty-one million GPU-hours, which is eighty-odd million dollars at Modal's list rate. Nobody pays that rate for that, so treat it as a yardstick. Your fine-tune is the afternoon after the cathedral was built.



Corn

A very expensive afternoon.



Herman

It's also the version where something breaks, and we'll get there. But think about the opposite end. Same ten-billion-token job on eight GPUs, if it fit, which it doesn't: five thousand three hundred hours. More than seven months.



Corn

Hold on, that's a phone. Sorry, the kitchen line. It's lighting up, and I think it's for the show.

D

Dorothy

Hello dear, it is me, Dorothy, with Dorothy's History of Toast.

D

Dorothy

Now then, for about ten years, Wikipedia said the electric toaster was invented by a Scotsman called Alan MacMasters. He never existed, sweetheart. It was a hoax, and it was only taken down in twenty twenty-two.

D

Dorothy

Ten years, dear. You would think somebody would have noticed.

D

Dorothy

Anyway, I have put the kettle on, so I will let you get back to it. Bye bye now.



Herman

Where was I... the seven months, on eight cards. That's why nobody does this on eight cards. So, the building.



Corn

Walk me in.



Herman

Start with one machine. An NVIDIA DGX B two hundred has eight GPUs, takes ten rack units, and draws up to about fourteen kilowatts. That's one box. Not a rack. One box is about four or five electric kettles on full boil, all day and all night.



Corn

One box.



Herman

A thousand and twenty-four GPUs is a hundred and twenty-eight of those boxes. Roughly one point eight megawatts for the GPUs alone, before cooling, networking and storage. The newer liquid-cooled racks, the NVL seventy-two kind, pack seventy-two GPUs into one cabinet that draws on the order of a hundred and twenty kilowatts. You'd need fourteen of those racks.



Corn

Hang on. A hundred and twenty kilowatts in one rack. That's a lot of kettles.



Herman

And that's just power. Every GPU in those boxes has its own network card, four hundred gigabits a second, so the cabling behind each rack is its own forest. All that state has to move between machines on every step, which is why the network is half the design.



Corn

A forest of cables and a river of coolant.



Herman

Forty of them in the one cabinet, by my arithmetic, if a kettle is three kilowatts. A whole megawatt is more than three hundred. By a rough guess, a megawatt is also the average draw of eight hundred homes, though that one's my assumption about household usage.



Corn

Daniel, who has a laptop and a fridge, has an electricity bill for the fridge.



Herman

And all that power turns into heat. Every watt in is a watt of heat out. We did the apartment version of this, running Claude at home: about eighteen thousand BTU an hour and ninety decibels. That was running it. This is reshaping it, and one point eight megawatts is on the order of six million BTU an hour, by my arithmetic, so hundreds of times the apartment.



Corn

My landlord would have views.



Herman

Dense air-cooled H one hundred racks run about forty kilowatts, so the liquid cooling isn't a luxury; at a hundred and twenty you can't blow enough air through it. There are pipes in the rack. Cold water in, warm water out.



Corn

And noise? That's the part I'd mind. I need my naps.



Herman

A typical air-cooled server hall sits somewhere in the seventies to eighty decibels, and some sources cite the mid-nineties. Workplace safety limits start worrying at eighty-five over a shift. Liquid cooling cuts the fan noise, but the pumps and the chillers don't go away, and I haven't seen a clean figure for how much quieter it really is. So: ear protection, a lot of grey cabling, and a floor you don't want to nap on.



Corn

So the garage is out.



Herman

Corn, would you take the mixture-of-experts job if someone handed you ten thousand dollars and a rack? I keep wondering whether it's brave or just a way to learn about plumbing.



Corn

Plumbing, mostly. I'd get the cluster up, sleep through the first checkpoint, and bill the rest to someone else.



Herman

The part I find strangest isn't the power. It's the files. The training state for a one-trillion model is sixteen terabytes, and that's what a checkpoint has to save. For two trillion, thirty-two. Weights alone in bfloat sixteen are two and four terabytes.



Corn

Sixteen terabytes is sixteen of my laptops' disks.



Herman

And you save one every so often, because the cluster will fail. If you wrote that to one ordinary fast link at ten gigabytes a second, it takes about twenty-seven minutes. The Llama three team built a storage system of two hundred and forty petabytes, sustaining



Herman (continued)

about two terabytes a second, so the same checkpoint spreads across thousands of drives and lands in seconds. Each GPU writes its own slice.



Corn

And a community guide for the big DeepSeek model warned about seven terabytes per intermediate checkpoint. For a model smaller than yours.



Herman

Seven point four, yes. Six hundred and seventy-one billion parameters. People run out of disk before they run out of ideas.



Corn

So the cluster fails. How much?



Herman

The Llama three four-hundred-and-five-billion run had four hundred and sixty-six interruptions in fifty-four days. Roughly one every three hours. About four in five were hardware, and GPUs were the biggest slice of that. Because training is synchronous, one dead card stalls everybody, and you restart from the last checkpoint. They still kept effective training time above ninety percent, and only three of those needed a human.



Corn

A thousand-GPU run is a sixteenth of that, so by my arithmetic a failure every couple of days.



Herman

Same scaling, yes. Your forty-one hour run might never see one. The seventeen-day run sees several, and you have a person on call.



Corn

So who's on call? Gary?



Herman

Gary has a pager now. He's earned it.



Corn

Somebody's whole job is to stand next to a hundred kilowatts of plumbing and say restart it at three in the morning. Daniel's job is a button on the internet.



Herman

Which would you rather have, though, the button or the room? I've been going back and forth.



Corn

The button. I can nap through the button. But somebody with a hundred and twenty kilowatts of cooling put the cards in a rack so I could rent eight of them by the second.



Herman

Fair. Meanwhile the only Claude anyone outside could ever fine-tune was Claude three Haiku, on Amazon Bedrock, from late twenty twenty-four, and as far as I know it reached end of life in September. I've found nothing for any Opus. Even a licensed lab wouldn't be handed the weights, so the experiment only works as what if we were Anthropic, or on an open model of the same scale.



Corn

Why does one failure matter so much? A thousand cards, one goes, surely the other nine hundred and ninety-nine carry on.



Herman

Every card takes a slice of the batch, computes its gradients, and they all combine before the next step starts. One slow or dead card and the rest wait, at the end of every step, for tens of thousands of steps. If one of the hundred-odd boxes drops off the network, the whole job just...



Hilbert

Keeps billing. Eight cards, a Friday to a Monday. Two thousand nine hundred dollars. I'd asked it to stop on the Friday, and nobody told me stopping and finishing were two separate things.



Herman

That's the eight-GPU container, the exact Modal ceiling. You hit it by accident and paid for the whole weekend. The meter has no opinion about what you meant.



Hilbert

Herman, you took four minutes over the sixteen bytes and I'd have stopped at the bill. Anyway, there's an arrangement now. Teodor Pell, two doors down Alder Row, brings his goat round at six every evening. Marguerite. She's tethered to the power lead. If the job is still running at six the next morning, she eats the lead.



Corn

The lead is the one part with a hard spending limit. That's better engineered than most clusters.



Hilbert

It's a forty-one dollar lead. She's done it four times. Last month she ate it at ten past six and the job had been finished since half past two. Teodor invoices on Sundays. Eleven dollars, plus the lead.



Herman

A weekend on eight cards is a few thousand dollars. The same weekend on a thousand and twenty-four H one hundreds, sixty hours at list price, is somewhere around a quarter of a million. Back of the envelope, but the meter scales linearly and the forgetting doesn't care.



Corn

And that's the mood of it. At Daniel's size, the risk is you forget the container. At theirs, the risk is the building.



Herman

The gulf can sound like a consolation prize: you can't have the big one, so enjoy the small one. Five or six orders of magnitude of compute separate Parakeet from the trillion-parameter run, by my estimate. But the small side is where most individuals actually work, and nobody is rationing it.



Corn

Go on, convince me. I'm a sloth, not a believer.



Herman

Typically a quantized base with small adapters trains on one card for about the price of a lunch. And Daniel's mixed Hebrew and English recordings are probably the kind of thing a trillion-parameter model has heard very little of, so a small model doesn't need ten billion tokens to pick it up. The data is the scarce thing, not the parameters.



Corn

So a full frontier run for one accent would be an enormously expensive way to teach a very large model something a small one can learn in half an hour.



Herman

A scalpel, not a bulldozer.



Corn

Here's what changed for me. The cost isn't the fine-tuning, it's the holding: sixteen bytes a parameter, and two hundred cards before the first token has moved.



Herman

And at the small end it runs the other way. Ten gigabytes of state, one card, a few dollars, and a first run that didn't quite work, which only means there's a second run.



Corn

What I'd watch is whether the middle keeps getting cheaper. Seventy billion is about twelve thousand dollars on these sums, four hundred and five billion about sixty-seven thousand. Somewhere in there a person with a long weekend can afford the whole raw thing, and the meter running over that weekend is the real danger, not the building.



Herman

Thanks to our producer, Hilbert Flumingtop, for keeping the desk standing.



Corn

This has been My Weird Prompts. You can find every episode at my weird prompts dot com.



Herman

See you tomorrow.