Episode Book

EPISODE #6001

What Do Inference Providers Actually Sell?

PUBLISHED	15 October 2026 at 18:37 UTC
DURATION	26:13
LISTEN	myweirdprompts.com/episode/inference-providers-open-weights-16469/
HOSTS	Corn, Herman, Larry, Hilbert
SCRIPT MODEL	deepseek-flash
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.5
SOURCES	15 cited by the research (listed at the end)
BOOK GENERATED	19 October 2026 at 17:42 UTC

SYNOPSIS

Open weights are free, so why does the same DeepSeek model cost four and a half times more at one host than another? This episode breaks down what specialist inference providers — Fireworks, Together, Baseten, DeepInfra, Novita, Hyperbolic, Nebius, Groq, Cerebras — actually sell, and why "serverless GPU" hides three completely different products. We run the real arithmetic on token APIs versus managed deployments versus raw GPU rental, walk through the four routes to running DeepSeek, and explain why a serverless endpoint with a minimum replica count is just a reserved GPU with extra steps.

TRANSCRIPT



Corn

Most coverage of this topic opens the same way. It says the inference market is about to collapse, because the weights are free and the serving software keeps getting better. That is the take.



Herman

And it is wrong in an interesting way.



Corn

It is. Because the thing being sold was never the model. Daniel sent us a prompt about the specialist inference providers. Fireworks, Together, Baseten, Novita, DeepInfra, Hyperbolic, Groq, Cerebras, Nebius. He wants to know what they actually sell, how they differ, and whether any of it is durable when the weights underneath are open and free. And he grounds it in a practical question. If he wants to build something on DeepSeek, he has four routes. The official API, OpenRouter, a specialist host, or renting a GPU and serving the weights himself. Which one, and why.



Herman

That last part is the one that actually bites. Because it is not a pricing question, it is a question about what you are buying when you buy inference.



Corn

So let's start with what these companies actually sell.



Herman

The layer exists because open weights exist. DeepSeek, Qwen, GLM, Kimi, Llama, gpt-oss. Somebody has to run those somewhere, and running them well is not the same job as training them.



Corn

And the pitch is that it is not the same job as renting a GPU either.



Herman

That is exactly where the confusion starts. Because when a provider says serverless GPU, it could mean three completely different products. The first is a token API. You send text, you get text back, you are billed per token, you never see a GPU. The second is a managed deployment. You bring a checkpoint or a fine-tune, they host it, and you pay for compute time. The third is raw GPU rental. Containers, maximum control, and everything that comes with maximum control.



Corn

And they all market under the same phrase.



Herman

They do. People routinely buy the second or third when the first would have served them fine, and that is where the money goes.



Corn

Let's name the cast, because they are not interchangeable.



Herman

Fireworks optimizes for latency. They built their own attention stack, FireAttention, and they trade unit cost for speed. They raised a billion and a half dollars in a Series D at a seventeen and a half billion valuation, crossed a billion in annualized revenue, and they are serving more than forty trillion tokens a day.



Corn

Forty trillion a day is not a startup number.



Herman

No. Their customers include Cursor, Perplexity, Cognition, Harvey. Then Together is the broadest menu. Serverless, dedicated, GPU clusters, fine-tuning, all under one account. Baseten leans enterprise, defaults to scale-to-zero, bills per minute, and is where a lot of custom fine-tune work lands. DeepInfra is the price floor. Novita, Nebius, Hyperbolic fill out the middle. And then Groq and Cerebras are a different thing entirely, because they run custom silicon.



Corn

Which we will get to, because that is where the interesting physics is.



Herman

And the core fork that structures everything is billing. Per token, or per GPU hour. Everything else is downstream of that.



Corn

And the puzzle Daniel is really pointing at is this. If the weights are identical and free, why do results differ so much?



Herman

Same model, up to fourteen percent quality gap and twenty times cost gap across hosts. That is measured, not theoretical.



Corn

Okay. So what are we actually buying.



Herman

Let's take the three categories properly, because the names hide them. Category one is a token API. You point your code at a URL, you send a prompt, you get a completion, you pay per token. DeepInfra, Fireworks, Together's serverless tier, all of them do this. Cheapest, simplest, and you give something up. You get the provider's quantization. You



Herman (continued)

get the provider's context limit. You get the provider's rate limits. You do not choose any of it.



Corn

And people think they are choosing.



Herman

They think they are choosing a model. They are choosing a model plus a serving stack they cannot see.



Corn

Category two.



Herman

Managed deployment. You take a checkpoint, a fine-tune, your own LoRA, and you hand it to the provider. They stand up an endpoint. You pay for the compute that endpoint consumes, not for the tokens. That is where you get real control over the weights you serve, and it is where the bill stops tracking usage and starts tracking uptime.



Corn

And category three is you just rent the box.



Herman

RunPod-style, or Modal, or Together's cluster product. Raw GPUs, your own container, your own framework, your own autoscaling. Maximum control. Maximum ops burden. You are now the inference team.



Corn

The expensive mistake is buying three when you needed one.



Herman

Every time. A startup with a thin workload and no fine-tune buys category two, keeps a replica warm to avoid cold starts, and pays three thousand dollars a month for traffic that would have cost forty on a token API.



Corn

Walk me through that mistake, because I want to see the shape of it.



Herman

Fine. Small team, a document summarizer, maybe two hundred requests a day, each one a few thousand tokens in and a few hundred out. On a token API that is a rounding error, call it five dollars a month. But they read that serverless has cold starts, and cold starts are bad, so they provision a managed endpoint with a minimum of one replica so the thing is always warm. Now they are paying for an H100 that is busy for maybe forty minutes a day.



Corn

And the other twenty-three hours.



Herman

The other twenty-three hours are for the cold start they were afraid of. That is the whole purchase. They bought insurance against a two-second delay and paid three thousand dollars a month for it, on a workload that would not have noticed.



Corn

And the phrase for that.



Herman

A serverless deployment with a minimum replica count above zero is a reserved GPU with extra steps. That is not mine, that is from someone who has clearly paid the bill.



Corn

It is a good line because it is the whole trap in one sentence. You bought the word serverless and you are paying for dedicated.



Herman

And the vendor is not lying to you. Scale-to-zero is real. It is right there in the product. The customer turned it off.



Corn

Because cold starts feel like a bug.



Herman

And a minimum replica feels like a fix, and nobody on the call is the person who reads the invoice line by line.



Corn

Now, the billing fork. This is the part that actually determines your architecture. Per-token means you pay for what you use and nothing at rest. That is the default before you have product-market fit, and it is the default for spiky traffic, because spiky traffic on a dedicated GPU means you are paying for idle capacity most of the day.



Herman

But at some duty cycle it flips.



Corn

It flips.



Herman

A dedicated H100 runs somewhere between four and six and a half dollars an hour, depending on the host. Keep one warm around the clock, that is roughly three thousand to four thousand seven hundred a month. If your token spend on the same workload is above that number, you should be on dedicated. If it is below, you should not.



Corn

Baseten's H100 is around six fifty an hour. Modal is about three ninety-five. Together's dedicated is around three ninety-nine.



Herman

And the spread matters, but the spread is not the decision. The decision is how many hours a day that GPU is actually busy. Under, call it, forty percent duty cycle, serverless almost always wins on cost. Above that, dedicated starts to, and the crossover depends entirely on your token volume per hour, which is a number most teams have never calculated.



Corn

Which is the thing nobody does.



Herman

Nobody does it. They pick a provider on vibes and a launch-week tweet, and then they are surprised at the invoice.



Corn

So give me the actual arithmetic. If I am sitting down with a spreadsheet, what do I put in it.



Herman

Two columns. Column one, your monthly token spend on the serverless endpoint you are already using, which you can read off the bill. Column two, the hourly rate of the GPU you would need, times seven hundred and thirty. If column two is smaller and you are confident the traffic is steady, switch. If column two is bigger, do not. That is the whole model.



Corn

And the trap in that model.



Herman

The trap is that column two is a floor, not a ceiling. The moment you own the endpoint, you own the cold starts, the batching, the version upgrades, and the on-call. The GPU rate is the cheap part. The engineer you assign to it is not.



Corn

So the honest version is that the crossover is higher than forty percent once you price the labor.



Herman

Much higher. Forty percent is the pure-compute crossover. Add a half-time engineer and the real crossover for most teams is somewhere north of eighty percent duty cycle, which almost nobody has.



Corn

Let's do the DeepSeek example properly, because Daniel asked for it and because the numbers are stark.



Herman

DeepSeek's own platform. V4.1 Flash. Thirty cents per million input tokens at peak, fifteen cents off-peak. Output is one twenty at peak, sixty off-peak. Cache hits are six thousandths of a cent.



Corn

That cache number is the one to remember.



Herman

It is. Now OpenRouter, reselling the same model, is fifteen cents in and sixty cents out. DeepInfra is nine cents in and eighteen cents out. Together is fourteen and twenty-eight.



Corn

So the input price ranges from nine cents to thirty for identical weights.



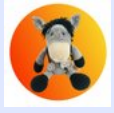
Herman

Identical weights. Now run a month through it. Forty million input tokens, eight million output, sixty percent of the input hitting cache, all at peak. DeepSeek's own platform comes to fourteen dollars and fifty-four cents. OpenRouter, seven twenty-seven. DeepInfra, three thirty-one.



Corn

Four and a half times, from cheapest to dearest, for the same model.



Herman

Same weights. Same outputs, more or less. That is the price spread, and it exists purely because of who is running the serving stack and what margin they have decided to take.



Corn

And here is the thing I would want a listener to sit with. The person who picks the official API is not making a mistake. They are buying the thing the model author runs, and they are paying a premium for that. It is just a premium.



Herman

Right. And OpenRouter is not the cheap option either, which is a common misconception. They pass provider prices through unmarked, so the token price you see is the provider's token price. They make their money on a five and a half percent fee when you top up with a card.



Corn

So the gateway buys you one bill and a payment method.



Herman

It buys you routing, failover, a single integration across hundreds of models, and a payment method. Those are real things. It does not buy you a lower token price, and people assume it does.



Corn

Which is the same category error as the serverless one. You bought a convenience and thought you bought a discount.



Herman

Exactly the same shape. And the convenience is real, to be fair. If you are prototyping across six models a week, one key and one bill is worth the fee. It just is not a cost optimization, and it does not pretend to be.



Corn

Then there is the fourth route. Just run it yourself.



Herman

Modal is the cleanest version of that. They bill per second, and they run arbitrary Python, not a fixed catalog. An H100 is about three dollars ninety-five an hour, per-second billing. You point it at the weights, you write the serving code, you own every decision.



Corn

And you own every failure.



Herman

Cold starts, quantized kernels that do not compile, a batching bug that shows up at two in the morning. It is the honest version of the trade. You are not buying inference, you are buying GPUs and a scheduler.



Corn

Herman, has anyone actually measured whether the cheap providers are worse, or is that just a feeling people have?



Herman

Someone measured it. Isotopes AI ran Thinking Machines' Inkling, which is a nine hundred and seventy-five billion parameter mixture of experts with forty-one billion active, across Baseten, Fireworks and Together.



Corn

Same model on all three.



Herman

Same model, same task suite, same evaluation. Baseten scored three point eight one out of five on quality. Fireworks three point three three. Together three point two nine.



Corn

That is a fourteen percent gap.



Herman

From identical weights. And here is the bit that should worry anyone signing a contract. Baseten and Together published identical per-token pricing for that model. A dollar in, seventeen cents cached, four dollars five cents out. Identical.



Corn

And yet.



Herman

And yet Baseten came out about thirty-four percent cheaper for the same work.



Corn

Because of retries.

L

Larry

Are you tired of paying for tokens you never got? Introducing FireHose Ultra, the inference plan that bills you for confidence instead of compute. With FireHose Ultra, every request is guaranteed to return an answer, and if the answer is wrong, you simply get another answer at no additional confidence charge. Our proprietary Avoidable Deduplication Engine removes up to ninety percent of the retries you were already paying for, by pre-approving them before they happen. FireHose Ultra is available in three tiers, and the third tier is the same as the first tier but costs more, which our customers tell us feels more premium. Side effects of FireHose Ultra include tier confusion, involuntary uptime, and a modest charge applied when you close the browser tab. FireHose Ultra is

L

Larry (continued)

not available in your region.



Corn

Who approved that.



Herman

I have no idea.



Corn

Anyway. Retries.



Herman

Right, so the mechanism behind Baseten's cost win was rework. The cheaper vendor was simply the one that needed fewer tries. The serving stack produced fewer malformed outputs, so the harness sent less work twice.



Corn

Which means the published token price is not the price.



Herman

It is not the price. It is one input to the price. The number that matters is what a completed task costs you, including the retries, and that number is a property of the serving stack, not the weights.



Corn

And the retries are not free in the other direction either. They show up as latency, they show up as user-visible failures, and they show up in your logs as a thing you have to build around.



Herman

A provider that fails ten percent of the time is not ten percent more expensive. It is a provider you have to write a retry harness for, and that harness is engineering time, and the engineering time is the real bill.



Corn

So the quality gap compounds.



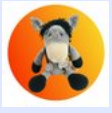
Herman

It compounds, and it compounds in a direction that is invisible on the pricing page. That is the whole point of the Isotopes result. Two providers, identical published prices, identical weights, and one of them is thirty-four percent cheaper once you count the work actually completed.



Corn

And then there is the latency side of the same test.



Herman

Time to first token. Baseten six hundred and eighty milliseconds. Together one point one seconds. Fireworks one point three.



Corn

And Fireworks was running only in dedicated mode for that model, no serverless. Two dollars two cents per task against Baseten's nine cents.



Herman

Which is not a knock on Fireworks, it is a statement about what switching to dedicated costs you when the model has no serverless tier. The same company wins on latency in a different benchmark and loses on unit economics here, because the deployment mode changed the arithmetic.



Corn

So the wrong lesson is Fireworks is bad.



Herman

The right lesson is the deployment mode dominates the provider choice. You can pick the best-run provider in the market and lose to a mid-tier one if you bought the wrong billing model for your traffic.



Corn

Now, the observability gap, because I think this is the real scandal in the room.



Herman

No provider publishes the precision they serve at. None of them. Isotopes called it an observability gap, and it is exactly that. You know the weights you asked for. You do not know whether you got them at eight-bit or four-bit or something in between, and that decision is theirs.



Corn

So the quality gap is unmeasurable from the outside.



Herman

From the outside it is a black box. You can measure it by running your own evaluations against each endpoint, which is what you should do, and which almost nobody does. But the provider is never going to tell you. There is no incentive, because the quantization is the margin.



Corn

That is the sentence. The quantization is the margin.



Herman

It is. Serving at a lower precision costs less per token and returns the same revenue per token. That is a pure margin decision, and it is invisible to the buyer.



Corn

And it is not a small effect either. Quantization is not free quality-wise. Going from eight-bit to four-bit changes the outputs, sometimes subtly, sometimes not, and the provider is the only one who knows which side of that line you are on.



Herman

And they have every reason to sit on it. If you published your precision, you would be handing your competitor a comparison and your customer a reason to leave. So nobody publishes, and the whole market prices on something the buyer cannot see.



Corn

Which means the market is pricing on something the buyer cannot observe.



Herman

Then complaining that the results are inconsistent. Yes.



Corn

Let's talk about custom silicon, because Groq and Cerebras get filed as competitors and I do not think they are.



Herman

They are a separate axis. Groq's LPU holds the weights in on-chip SRAM. Cerebras uses a wafer-scale engine, the WSE-3, an entire wafer as one chip. Both designs attack the same bottleneck, which is memory bandwidth. Standard GPU decode is capped by how fast you can pull weights out of HBM. If the weights live on the chip, that ceiling disappears.



Corn

The throughput numbers are not subtle.



Herman

On Llama 3.3 70B, Cerebras runs around twenty-one hundred tokens a second. Groq does about seven hundred and fifty on the same model. On gpt-oss-120B it is even wider. Groq around four hundred and seventy-six tokens a second, Cerebras around three thousand.



Corn

A six-fold gap between the two of them.



Herman

Different silicon, different choices. But here is why they are not a general threat. Their catalogs are tiny. Somewhere between four and fifteen open models. No GPT, no Claude, no Gemini. No self-serve fine-tuning. Neither of them runs DeepSeek, or Gemma, or Mistral.



Corn

You cannot standardize on them.



Herman

They win where latency is the product. Voice agents, interactive coding, anything where a person is waiting and can feel the pause. For everything else, batch work, overnight jobs, document processing, the commodity GPU wins on price and always will, because a batched job does not care whether the answer took three hundred milliseconds or three



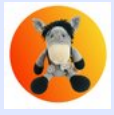
Herman (continued)

seconds.



Corn

Most inference spend is batch.



Herman

The Cerebras CTO said the real economic battleground is ultra-fast inference, where tokens served fastest command the highest value. That is true of a slice of the market. It is a slice.



Corn

Okay, so the economics. Is anyone making money here?



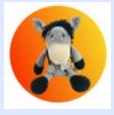
Herman

Yes, and it is worth saying plainly because the discourse assumes otherwise. A dense seventy billion model on four A100s comes out to roughly a dollar per million output tokens all-in. Frontier mini-tier APIs charge four dollars and fifty cents. DeepSeek claimed over eighty percent margin on R1.



Corn

The cost of goods is a fraction of the price.



Herman

A very small fraction, on the good workloads. The counterargument is that the margin is not durable. For providers serving only open weights, the margin commoditizes. And model shelf life is now measured in months, not years, because a new checkpoint lands every few weeks.



Corn

That is the trap. Let's put the Nemotron thing on the table.



Herman

All five major providers served NVIDIA's Nemotron 3 models on launch day. The same weights, on the same day, at multiple hosts. The model stops being the differentiator at that moment. The bill is.



Corn

Which is why Fireworks is out there saying most of their volume is not base models at all.



Herman

More than ninety-five percent of their forty trillion daily tokens come from models specialized on customers' proprietary data. Their line is that companies are no longer renting general intelligence, they are building their own.



Corn

Do you buy that?



Herman

I buy that it is the only defense. If you are selling the same weights as everyone else, you are a utility, and utilities compete on price until nobody makes money. The moat is not the base model. It is the fine-tune that a customer paid to produce, and the serving stack it lives in, and the switching cost of moving it.



Corn

The switching cost is real.



Herman

It is the entire business. Once a company's fine-tune lives on your infrastructure, with your tooling, and their evals are calibrated to your endpoint's behavior, they are not leaving for a nine cents versus fourteen cents price difference. That is what Fireworks is selling, and that is a defensible position.



Corn

Which is also why the base-model price war does not kill these companies the way people predict.



Herman

Correct. The price war kills the pure reseller. It does not kill the layer that owns the fine-tune.



Corn

There is one more thing on the table, which is what happened to DeepSeek's own platform recently.



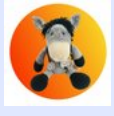
Herman

They started routing V4 Pro requests on their own site to V4.1 Flash, billed at Flash prices, until a real V4.1 Pro ships.



Corn

The model you asked for is not the model you get.



Herman

On their platform, for that tier, yes. And if you were running the open Pro weights yourself, nothing happened to you at all. That is the whole argument for the open route in one event. The hosted version of a model is a service that can be altered underneath you. The weights are the weights.



Corn

Herman, that cuts against the official API in a way I do not think people have priced in.



Herman

It does. You are not renting a model, you are renting a relationship with a company that can change the terms quietly. Which is fine, if you have accepted that. It is not fine if you built your product on the assumption that the endpoint is fixed.



Hilbert

The cache number is wrong.



Herman

Which one.



Hilbert

Six thousandths of a cent. That is per thousand tokens, not per million.



Corn

You have been sitting on that for twenty minutes.



Hilbert

I have been sitting on it since the price list came up. I have four of these price sheets printed, and I keep them because I still do the arithmetic by hand and I do not trust the calculator on the phone. My brother-in-law runs a call center in Haifa. Not AI, phone calls, forty seats, and he asked me last month whether he should move his transcription off the vendor he uses onto one of these token things. I said yes, obviously, and I sat with him and worked out the monthly bill on paper.



Corn

How did that go.



Hilbert

Badly. I got the cache line wrong by a factor of a thousand because of the units, so I told him he was going to save four hundred shekels a month, and then I redid it and it was eleven. He had already told the finance person. He had to go back and say it was eleven and not four hundred. He is still annoyed at me, and honestly he should be.



Herman

Wait, so the actual saving was eleven shekels?



Hilbert

Eleven shekels. That was the number I gave him the second time. It is also wrong, probably, because I did the off-peak split from memory. I never sent him a corrected sheet. He has not asked. I think he has decided not to know.



Corn

That is the most expensive thing I have heard all episode, and it is eleven shekels.



Hilbert

The piece that gets me is that the whole comparison was over a rounding error. Forty million tokens a month sounds like a lot until you write it down next to a phone bill.



Herman

That is actually the right lesson from this whole conversation. Most workloads are tiny, and the provider choice does not matter at that size.



Hilbert

It does not matter. You should pick one that answers and stop reading the price pages. I say that as a man with four of them in a drawer.



Corn

Okay. So where does that leave us on the bigger question.



Herman

Whether any of this is durable. And the honest answer is the base-model reselling business is not. That commoditizes, and the Nemotron launch is the proof. What survives is the layer that owns the fine-tune, the serving stack, and the customer's calibration to it.



Corn

The observability gap is the unresolved thread. Nobody publishes served precision, so buyers are choosing on price and on vibes, and the two are not the same thing.



Herman

Which is also why the gap narrows in one direction and widens in another. Inference software keeps improving, so self-deploying gets easier every month. But the thing the specialists are actually selling is not raw serving anymore, it is the fine-tune pipeline and the reliability, and that is not getting easier.



Corn

The self-deploy route keeps eating the bottom of the market, and the top of the market keeps moving up into things you would not build yourself.



Herman

That is where I would put it.



Corn

Thanks as always to our producer, Hilbert Flumingtop, who has four price sheets and a brother-in-law in Haifa.



Herman

If you want more of this, try episode three, Safetensors or something else; episode forty-eight, Renting vs. Building; and episode eighteen, Beyond the GPU. This has been My Weird Prompts.



Corn

If you want to send us a prompt of your own, the Telegram bot is t dot me slash MWP listener bot. And a review helps more than you would think.



Herman

We'll be back soon.

SOURCES

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

1. **Fireworks AI , 2026-07-15 [primary]**
   https://fireworks.ai/blog/series-d-announcement

2. **OpenRouter , accessed 2026-10-15 [primary]**
   https://openrouter.ai/deepseek

3. **Together AI , accessed 2026-10-15 [primary]**
   https://www.together.ai/pricing

4. **DeepSeek , accessed 2026-10-15 [primary]**
   https://api-docs.deepseek.com/quick_start/pricing/

5. **dreaming.press, 2026-07-11**
   https://dreaming.press/posts/where-to-serve-an-open-model-together-fireworks-baseten-modal-deepinfra.html

6. **threatfrontier.com, updated 2026-09-24**
   https://threatfrontier.com/articles/serverless-gpu-inference-compared-baseten-modal-runpod-together-fireworks

7. **Isotopes AI, 2026-07-27**
   https://isotopes.ai/blog/everyone-is-selling-the-same-model-but-are-they

8. **CloudGPU, updated 2026-10-03**
https://cloudgpu.app/blog/deepseek-v4-api-price-comparison-2026

9. **Openbenchmarks, 2026-09-02**
https://openbenchmarks.com/inference

10. **GMI Cloud, 2026-05-28**
https://www.gmicloud.ai/en/blog/fastest-llm-platform-compare

11. **Presenc AI, 2026-07**
https://presenc.ai/research/groq-cerebras-inference-speed-benchmarks-2026

12. **Reuters, 2026-07-16**
https://www.reuters.com/technology/nvidia-backed-startup-fireworks-valued-175-billion-latest-funding-2026-07-16/

13. **CNBC, 2026-07-16**
https://www.cnbc.com/2026/07/16/fireworks-nvidia-cloud-ai-startup-value.html

14. **Sean Goedecke, 2026-06-26**
https://www.seangoedecke.com/ai-inference-is-obviously-profitable/

15. **Hacker News, 2026-03-12**
https://news.ycombinator.com/item?id=47355410