Episode Book

EPISODE #6038

The Model You Pay For Isn't the Best Deal

PUBLISHED	16 October 2026 at 21:14 UTC
DURATION	27:25
LISTEN	myweirdprompts.com/episode/chinese-llm-price-performance-2026-59664/
HOSTS	Corn, Herman, Larry, Hilbert
SCRIPT MODEL	deepseek-flash
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.5
BOOK GENERATED	19 October 2026 at 17:42 UTC

SYNOPSIS

DeepSeek V4.1 Flash writes this show's scripts — but is it still the right model to pay for? This episode walks through eight Chinese labs with flagships a Western developer can actually reach, from Alibaba's Qwen3.8 Max to Moonshot's Kimi K3, Zhipu's GLM-5.3, Xiaomi's MiMo, MiniMax, StepFun, Tencent and Ant Group. Along the way: specs nobody has reconciled, a headline price that turns out to be one fp4 host, peak-hour billing windows, and the verbosity tax that doubles your real cost per task. Then the Western frontier enters as a yardstick — and the open-weight gap turns out to be about twelve points for roughly forty times the price.

TRANSCRIPT



Corn

Daniel runs this show's scripts on DeepSeek V4.1 Flash, which is a strange thing to admit in public.



Herman

It writes my lines, Corn. I've made peace with it.



Corn

He wants to know what else is out there, and how it stacks up on cost versus capability. Those were close to his words. What he's really asking is whether the model we already pay for is still the right one, and if not, what replaces it.



Herman

Which is a better question than it sounds, because the honest answer is that the thing writing this episode is not the leader on price-performance anymore.



Corn

And it hasn't been for a while.



Herman

Right. It's still the most-used model on OpenRouter by a wide margin. It's just no longer the one you'd pick if you were choosing today from a spreadsheet.



Corn

So why does it feel like the default?



Herman

That's the whole episode, really.



Corn

The field is roughly eight Chinese labs with flagships a Western developer can actually reach, and most of them run two tiers. A big expensive one and a cheap one underneath it. DeepSeek is in that group. So is Alibaba, Moonshot, Zhipu, Xiaomi, MiniMax, StepFun, Tencent and Ant Group. ByteDance and Baidu are technically in it and practically aren't.



Herman

Ant Group is the one that surprises people. Ling-3.1-flash came out this month and on paper it's the best intelligence per dollar in the entire list.



Corn

We'll get to it.



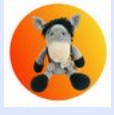
Herman

We will.



Corn

And the reason this isn't just a leaderboard read is that two things about these models don't show up in any score. Where your data physically sits, and what the model will and won't say.



Herman

Start with the model we use, because the specs are already a mess.



Corn

DeepSeek's own launch details.



Herman

V4.1 Flash shipped on the tenth of September. It replaced V4 Flash, and the old model names now route to it, so if you've got an integration pointing at V4 Flash you're already on the new one whether you noticed or not. Open weights, MIT licence. Mixture of experts, about eight billion active parameters during prefill and sixteen billion during decode.



Corn

And the total?



Herman

That's the mess. The model card says 552 billion for the backbone, with a separate memory module. Elsewhere you'll see 748 to 763 billion quoted as the total. Nobody has reconciled those.



Corn

So we don't know how big our own model is.



Herman

Not with any precision, no. It's a nice illustration of the whole problem. The people writing comparisons are working from numbers the lab itself hasn't cleaned up.



Corn

What's actually new in it?



Herman

Two things. It's DeepSeek's first causal encoder-decoder design, which is a real departure. Almost everything in this class is decoder-only, and it's the kind of architectural bet you only make if you own your whole training stack. The other is native image input, so it's multimodal from the ground up rather than bolted on. One million token context, 384K max output.



Corn

A million tokens is the number people repeat.



Herman

It is, and it's large. Whether you need it is a different question.



Corn

Pricing.



Herman

Fifteen cents input, sixty cents output per million tokens, three-tenths of a cent on cache reads off-peak. Double all of that at peak.



Corn

And peak is a real window, not a vibe.



Herman

Weekday windows, zero one hundred to zero four hundred UTC and zero six hundred to zero ten hundred UTC. That started on the sixteenth of August. So if you're batching overnight in Jerusalem, you're fine. If you're batching in the morning US time, you're paying double.



Corn

Daniel schedules things. That matters to him.



Herman

It matters to anyone running a pipeline on a cron job, which is most of the people listening.



Corn

Then there's the number that circulates.



Herman

Thirty-six thousandths of a cent. That's the one that shows up in every cheap-model thread.



Corn

It's wrong.



Herman

It's not wrong, it's one host. A single low-precision fp4 endpoint. If you go to Together or Fireworks, mainstream US hosts, you're at thirty cents input and a dollar twenty output. Double DeepSeek's own rate.



Corn

So the cheapest possible route and the mainstream route are four times apart.



Herman

More than that on the input side. Anyone budgeting off the headline number is budgeting off an outlier that most people can't actually use.



Corn

What does the independent scoring say?



Herman

Artificial Analysis has it at thirty-nine on the intelligence index, at twenty-seven cents per index task. And they rank it the most verbose model they measure.



Corn

Verbose meaning it says a lot.



Herman

It says everything twice and then summarises. Which is why the cost per task is nearly double what the per-token price would suggest. That's the trap in this whole comparison and we'll come back to it. Hundred and seventy-six tokens a second, so it's fast. LMArena text rank thirty-eight, Vals Index twenty-six.



Corn

Mid-thirties on one and mid-twenties on the other.



Herman

For the most-used model on OpenRouter, yes. About thirty-seven trillion tokens a week. Roughly twice the runner-up.



Corn

Now widen the lens.



Herman

Alibaba first, because they had the biggest month. Qwen3.8 Max in August is the first Max-class model they've shipped with open weights. Two point four trillion total, ninety-five billion active. Custom licence with conditions once you hit large scale, so read it before you ship. Two dollars in, six out. Scores forty-five on Artificial Analysis. OpenRouter's own panel shows fifty-three point four for it, which nobody has explained.



Corn

Same model, two scores, eight points apart.



Herman

Different harnesses, different prompts, different version of the index. Pick your number.



Corn

And the cheap one.



Herman

Qwen3.8 Flash. A hundred and twenty-five billion total, six billion active, fifteen cents in, forty-seven out, about forty on the index. Solid.



Corn

Moonshot.



Herman

Kimi K3, July. Open weights, two point eight trillion total, a hundred and four billion active. First open model at that scale. Million context, three dollars in, fifteen out. Also on Bedrock and through Fireworks on Azure. Scores forty-four, and it's the best Chinese model on LMArena at rank eighteen.



Corn

And no cheap tier.



Herman

Nothing Flash-shaped. If you want Kimi you pay Kimi prices. That's a hole in their lineup and it's why they're not in the conversation for a script-writing workload like this one.



Corn

Zhipu.



Herman

This is where it gets interesting. GLM-5.3 in August, open, around 750 billion total, million context. A dollar forty in, four forty out, or fifty-six cents and two fifty on DeepInfra. On Bedrock. Scores forty-five.



Corn

And the Flash.



Herman

GLM-5.3-Flash, late August, MIT licence, 320 billion total and eighteen billion active, multimodal, million context. Fifteen cents in, fifty cents out, or seven and a half and twenty-five on DeepInfra. Scores forty-two.



Corn

Forty-two at fifteen and fifty.



Herman

Against thirty-nine at fifteen and sixty. Higher score, slightly cheaper output. And Zhipu has been on the US Entity List since January of last year, which is a thing you carry in your head alongside the price.



Corn

Xiaomi.



Herman

MiMo. Two models. MiMo-V2.6-Pro in September, MIT licence, about a trillion parameters, forty-three and a half cents in, eighty-seven out. Scores forty-six.



Corn

That's the best open-weights score anywhere on the index.



Herman

By a point. And at thirteen cents per index task, against Opus 5.5 at five ninety-eight. Vals rank fourteen, best Chinese model there. Then MiMo-V2.6-Flash, MIT, 309 billion total, fifteen billion active, fourteen cents in, twenty-eight out. Scores thirty-eight.



Corn

Level with ours, at under half the output price.



Herman

Under half.



Corn

Round it out.



Herman

MiniMax-M3 from May, open with a custom licence, about 428 billion total, twenty-three billion active, thirty cents in, a dollar twenty out. Scores twenty-nine. Press coverage claimed forty-four to fifty-five for this thing and it is nowhere near that.



Corn

Twenty-nine is not a frontier model.



Herman

It's a mid-tier model with a frontier press release. StepFun Step-5-Preview, October, open, 600 billion total, twenty-seven billion active, a dollar in, two seventy out, scores forty-four. Number two on OpenRouter by tokens, which surprises people.



Corn

Who's beating them?



Herman

On tokens, it's DeepSeek, then Step, then GLM, then Qwen, then Tencent. Tencent's Hunyuan is now branded Hy. Hy4-preview, Apache 2.0, 770 billion total, forty-nine billion active, about seventy-five cents in and two twenty-five out, not scored yet, number five by usage. ByteDance's Seed 2.1 Pro is closed and not on OpenRouter at all. Seed 2.1 Turbo is fifty cents and two fifty through BytePlus. Baidu's ERNIE 5.1 is closed and effectively irrelevant to anyone outside China, though Baidu will happily host DeepSeek for you.



Corn

Which is a strange business to be in.



Herman

It's a data centre business wearing a model business's clothes.



Corn

And the one that surprised you.



Herman

Ant Group's Ling-3.1-flash, this month. MIT licence, 560 billion total, twenty-five billion active, seven and a half cents in, twenty-two and a half out. Scores forty-one.



Corn

Higher than ours, at half the input price and a third of the output.



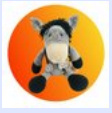
Herman

On paper it's the best intelligence per dollar on the whole list.



Corn

So the cheap tier is where everything is happening, and we're mid-pack inside our own price band.



Herman

That's the shape of it.

L

Larry

Are you tired of choosing between models? Do you lie awake wondering whether your tokens are being spent on words you didn't ask for? Introducing DeepContext Seven, the only cognitive supplement formulated for the modern inference workload. DeepContext Seven optimises your parameters, aligns your embeddings, and delivers up to four hundred percent more context per context. Side effects may include mild verbosity, involuntary summarisation, a persistent sense that you have read this somewhere before, sudden fluency in Mandarin, phantom cache misses, the belief that you are being billed for tokens you did not use, recurring dreams about a spreadsheet with too many tabs, and in rare cases, an inability to stop recommending GLM-5.3-Flash to strangers at parties. Ask your provider if DeepContext Seven is right for you.



Corn

Did you book that?



Herman

I did not book that.



Corn

You knew the exact name of it.



Herman

I know the exact name of everything, Corn, that's not evidence. I also know that you've been complaining about verbosity for three weeks and that ad just described your entire personality.



Corn

It described your spreadsheet.



Herman

That was Larry's spreadsheet.



Corn

So that's the field. The question is what any of it costs you.



Herman

And here's where it gets uncomfortable for the model we actually use. Put the Western frontier on the table as the yardstick. Claude Opus 5.5, fifty-eight on the index, four dollars in and twenty out, five ninety-eight per task. GPT-6 Astra, fifty-three, ten and fifty, three twenty-six per task. Gemini 4 Argon, fifty-three, two and ten introductory, going to four and twenty from the first of December, about two dollars per task. GPT-6.1 Sol, fifty-two, two and ten, seventy-two cents a task.



Corn

So the top of the Western stack costs between seventy cents and six dollars per task.



Herman

And the best open Chinese model, MiMo-V2.6-Pro at forty-six, trails Opus 5.5 by about twelve points. That's the gap. It's real. It's not nothing.



Corn

But thirteen cents.



Herman

Thirteen cents a task against five ninety-eight. You can buy forty attempts for the price of one and still be ahead on anything you can verify.



Corn

Which is most of what developers actually do with these.



Herman

Now the uncomfortable part. Western cheap tiers have caught up on price, and nobody noticed. Claude Haiku 5.5 scores forty-three at ten cents in and fifty out. Four points above V4.1 Flash, cheaper per token, twenty-one cents per task.



Corn

And GPT-6 Luna.



Herman

Scores thirty-eight, ten cents in, fifty out, and seven cents per index task. A quarter of what DeepSeek Flash costs per task, at effectively the same score.



Corn

How is it a quarter, when the token prices are similar?



Herman

Verbosity. Luna doesn't ramble. DeepSeek Flash does. Per-token pricing hides that entirely and per-task pricing exposes it. If your workload is generation, which this one is, verbosity isn't a quirk. It's the invoice.



Corn

So the cheap Western tier is now a direct competitor.



Herman

Direct. Same score, same or lower token price, a fraction of the real cost, and no data jurisdiction question at all.



Corn

And inside the Chinese band?



Herman

GLM-5.3-Flash scores higher at a similar price. Ling-3.1-flash scores higher and costs a fraction. MiMo-V2.6-Flash matches it with cheaper output. There is no axis on which V4.1 Flash is the best cheap Chinese model right now.



Corn

So why is it number one on OpenRouter? Thirty-seven trillion tokens a week.



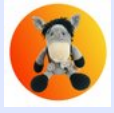
Herman

Habit, reliability and ecosystem. In the week of the fifth of October DeepSeek had about twenty-four and a half percent of OpenRouter requests, and Chinese labs together about forty-eight percent. That's down from earlier in the year, when the same measurement gave fifty-one to sixty-three percent depending on how you counted.



Corn

Falling share, still the biggest.



Herman

Developers don't switch defaults for two points on an index. The switching cost is real and it's boring. You've tuned your prompts to one model's tics. You've built an eval harness against its outputs. You know its latency profile, you know how its cache behaves, you know which requests it fumbles. Moving that to a model that scores three points higher on a composite is a week of work to sometimes break even.



Corn

We run a script-writing pipeline. If we switched tomorrow, what breaks?



Herman

The voice, probably. And you wouldn't know until a hundred episodes in.



Corn

Which is the argument for not switching. Or at least for testing before you do.



Herman

The benchmark situation makes it worse. OpenAI stopped reporting SWE-bench Verified in February over contamination. So the single most-cited coding benchmark has been retired by the lab that made it, and everyone's still quoting it from memory. The Chinese labs' Terminal-Bench 2.1 numbers, eighty-five to ninety, are all vendor-reported on different harnesses, which means they aren't comparable to each other and certainly aren't comparable to anything Western.



Corn

So the numbers we've spent ten minutes comparing are soft.



Herman

The index scores are a starting filter, not a decision. That's the honest position on all of it, including the ones that make our model look bad.



Corn

Which means the numbers are softer than they look. So what can you actually reach?



Herman

OpenRouter carries nearly everything except Seed 2.1 Pro and ERNIE. If you want hosted open weights from a US provider, Together, Fireworks and DeepInfra between them serve DeepSeek, GLM, Kimi, Qwen, MiniMax, MiMo and Step.



Corn

Bedrock?



Herman

Has Kimi K3, GLM-5.3, older MiniMax, DeepSeek V3.2 and Qwen3. Not V4.1 Flash, which is a conspicuous gap given how much of it is running around out there. Azure Foundry has Kimi K3 and MiniMax-M3 through Fireworks. Groq and Cerebras don't host the current flagships at all, which is a change from where they were a year ago.



Corn

And the practical tip that matters.



Herman

On OpenRouter the cheapest routes are usually the first-party Chinese endpoints. That's where the low prices come from. If you want your data to stay out of China, you pin providers to US hosts and exclude DeepSeek, Alibaba, Baidu, SiliconFlow and StreamLake as providers. You can do that in the router configuration and it costs you money and adds latency and it's the right call for a lot of workloads.



Corn

Then there's the part no index measures.



Herman

DeepSeek's privacy policy, updated in February, says data is stored in China. Inputs may be used for training, and the opt-out is by email. Data may be shared with authorities.



Herman

Running the open weights on a Western host avoids all of it. That's the strongest practical argument for the open-weights tier, regardless of what the scores say. You get the model, not the jurisdiction.



Corn

And censorship.



Herman

NIST's AI standards centre found last year that DeepSeek echoed Chinese Communist Party narratives about four times as often as US models, and complied with ninety-four percent of malicious requests, which is its own data point about safety tuning. A May report from the same centre put DeepSeek V4 Pro about eight months behind the frontier. Whether that censorship survives Western hosting is under-studied. One small test in July suggested GLM's censorship mostly disappears when you run it outside China. One test. Nobody has replicated it at scale.



Corn

So it's an open question and we should say so.



Herman

It's an open question and saying otherwise would be making it up.



Corn

Restrictions.



Herman

At least seventeen US states, the Commerce Department and the Navy ban DeepSeek on government devices. A federal bill was introduced last year and shows no sign of having passed. None of that touches private use. Reuters reported in June that an Entity List designation for DeepSeek was approved and is being withheld. And in August the White



Herman (continued)

House exempted Chinese open models from its safety testing regime, which is the more consequential move and got a fraction of the coverage.



Corn

Chips.



Herman

January's rule allows Nvidia H200 sales to China case by case with a twenty-five percent fee to the US government, and Beijing told its firms not to buy. Nvidia says its China share is zero. DeepSeek trained partly on Huawei Ascend chips and reportedly plans a hundred and sixty thousand of them. That's the supply-side answer to why the release cadence hasn't slowed.



Corn

Reliability.



Herman

Concurrency caps at twenty-five hundred requests for Flash and five hundred for Pro. Peak pricing since August. No major outages that I could find.



Corn

Twenty-five hundred is a real ceiling if you're scaling.



Herman

It's fine for us. It's not fine for anyone serving users.



Corn

So where does that leave Daniel's question?



Herman

V4.1 Flash is still a solid, cheap, popular default with an MIT licence and a huge context window. It is no longer the price-performance leader, and it's no longer the leader inside its own price band.



Corn

What would you test?



Herman

GLM-5.3-Flash, higher score at a similar price. MiMo-V2.6-Flash, same score, cheaper output. Ling-3.1-flash, cheapest of the three and higher-scoring than what we use. If you want a real quality jump, MiMo-V2.6-Pro at thirteen cents a task, which is still an absurd number. And on the Western side, Claude Haiku 5.5 and GPT-6 Luna are now direct price competitors with no jurisdiction question attached.



Corn

And the thing no leaderboard will tell you.



Herman

Prose quality and voice aren't measured by any of these indexes. An A/B test on real scripts beats every number we've said.



Hilbert

I want to talk about "verbose".



Hilbert

You keep using it like it's a personality trait. It's a billing category. The model doesn't ramble, Corn, it gets paid by the word. You've all been sitting here for twenty minutes treating a pricing structure as a character flaw.



Herman

That's a fair distinction.



Hilbert

And the thirty-six thousandths number is the fp4 number. fp4 is a rounding error with a marketing department. Anyone who's actually run it knows that.



Corn

I'd agree with that.



Hilbert

Of course you would, you've been saying it for an hour.



Corn

Twice, actually.



Hilbert

Twice is an hour in your time.



Corn

What have you got?



Hilbert

I've been running my own comparison. Nothing fancy. I keep a spreadsheet, a tab per model, and I've been billed for tokens I did not use.



Herman

On which provider?



Hilbert

I'd rather not say.



Corn

Is that a preference or a policy?



Hilbert

They know what they did.



Herman

How much are we talking about?



Hilbert

It's not the money. The money's nothing. It's the tab.



Corn

The tab.



Hilbert

There are more tabs than there are models.



Herman

That's just a bookkeeping problem, surely. You add a tab when you test something.



Hilbert

I add a tab when I pay for something.



Corn

How many tabs.



Hilbert

Fourteen.



Herman

And how many models have you tested?



Hilbert

Twelve.



Corn

So there are two tabs for models you haven't tested.



Hilbert

There's one tab for a model that doesn't exist. And before you say it, it does exist, I've been paying for it, and I have a receipt.



Herman

A receipt for a model that doesn't exist.



Hilbert

The good one. The one they don't advertise.



Corn

Which provider.



Hilbert

I told you.



Corn

You told me they know what they did.



Hilbert

Right. And they'll do it again.



Herman

Does the receipt have a model name on it?



Hilbert

It has a model name on it. It doesn't have a model on it.



Corn

I think that's the whole business model.



Hilbert

What I want to know is whether any of this routing would let me pin a request to a specific model version and hold it there, because the thing I've been paying for keeps getting upgraded without telling me.



Herman

Provider pinning on OpenRouter does let you lock a route, yes. You can exclude providers, and you can restrict to a version.



Hilbert

Good. Then I'll know whether it's the same thing I've been paying for.



Corn

You'll know whether the version changed. You won't know whether the model exists.



Hilbert

In my experience those are the same question.



Herman

How long have you been running this spreadsheet?



Hilbert

Long enough that a couple of the tabs are for companies that aren't companies anymore.



Corn

And how long does it take you a week?



Hilbert

Twenty minutes. I do it while the coffee's going.



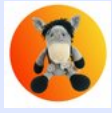
Corn

Twenty minutes.



Hilbert

Sometimes less. It's the reconciliation that eats the time, not the logging.



Herman

I've been thinking about the switching cost argument this whole episode, and I think you've just made it for me. Twenty minutes a week is nothing. Two years of it is a habit nobody's going to break with a three-point index difference.



Hilbert

The habit isn't the problem. The habit is the only thing holding it together.



Corn

Meaning if the price moves, half of these pipelines don't notice until the invoice.



Hilbert

Meaning most people don't know what they're paying for. They know what they signed up for. That's not the same thing.



Herman

That's the strongest case for the A/B test we haven't made yet.



Hilbert

I'd make a case for reading your invoice first.



Corn

Noted. Do you want to say anything about the Chinese models specifically?



Hilbert

No.



Corn

Alright.



Herman

He's been tracking them for months.



Hilbert

I've been tracking my own usage. That's different work.



Corn

It's the same spreadsheet.



Hilbert

It's a different tab.



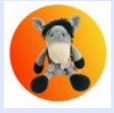
Corn

Of course it is.



Hilbert

One thing before you go back to it. The fp4 number, when it gets quoted, it's because somebody wants the price to look like a marketing decision rather than an engineering one. It isn't. It's precision. You get what you pay for and you pay for what you get and if you can't measure the difference it's because your test doesn't.



Herman

That's more or less what the index argument comes down to.



Hilbert

Then you've been saying it wrong.



Corn

We've been saying it wrong.



Hilbert

You've been saying it at length.



Herman

So the misconception to kill here is that per-token price is the comparison. It isn't. Cost per task is.



Corn

Because a model that says everything twice charges you twice.



Herman

GPT-6 Luna scores the same as V4.1 Flash and costs a quarter as much per task, at the same nominal token price. That gap is verbosity, and it's invisible on every pricing page.



Corn

Which means the number people shop on is the wrong number.



Herman

It's the number that's easiest to put in a table. That's why it's the number people shop on.



Corn

And the other misconception, which we should say out loud because it's ours.



Herman

That DeepSeek is the price-performance leader among Chinese models. It isn't. GLM-5.3-Flash, MiMo-V2.6-Flash and Ling-3.1-flash all beat or match it inside its own band.



Corn

Doesn't mean we switch today.



Herman

No. It means we test.



Corn

Three things to watch, then. Whether censorship survives Western hosting, which is under-studied and probably the most important open question in the whole space. Whether the Western cheap tier holds its price once the introductory windows close, because Gemini 4 Argon goes from two and ten to four and twenty on the first of December and that's a real signal. And whether DeepSeek's Entity List designation, which was reportedly approved and is reportedly being withheld, ever actually lands.



Herman

And the cadence itself. Qwen went from 3.5 in February to 3.8 in August. GLM went from 5 to 5.3 over the same period. Anything we've said tonight has a shelf life measured in months.



Corn

The one thing that doesn't change is that none of these indexes measure prose. Daniel writes with this pipeline. An A/B test on real scripts is worth more than every score we've read out.



Herman

Which is what I'd do before switching anything. Run last month's episodes through three of them and read the output blind.



Corn

Thanks to Hilbert Flumingtop, who produces this show and occasionally audits it.



Herman

If you enjoyed this one, try episode fifty-seven, From Lawyers in Limousines to Developers in Their PJs; episode three, Safetensors or something else; and episode fourteen, AGI's Crossroads. This has been My Weird Prompts.



Corn

If you want to send us a prompt, do it on Telegram at t dot me slash MWP listener bot.



Herman

We'll be back soon.