# Episode Book

EPISODE #6216

# What Do You Call the Middle Tier Model?

| | |
|---|---|
| PUBLISHED | 22 October 2026 at 23:25 UTC |
| DURATION | 19:28 |
| LISTEN | myweirdprompts.com/episode/middle-tier-ai-models-28027/ |
| HOSTS | Corn, Herman, 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

Every lab has a middle rung — Anthropic's Sonnet, OpenAI's Sol, Google's Flash, Mistral's Medium — but there's no agreed name for it, and the price lists have quietly converged into a ladder. This episode digs into why the tier most people use all day is getting squeezed by cheap open-weight models below and flagship promotions above. It's a conversation about cost per finished task rather than cost per token, why effort settings can matter more than the tier name, and what "the middle" even means when last year's flagship keeps becoming this year's workhorse.

# TRANSCRIPT



### Corn

The model most people reach for all day is getting squeezed from both directions, and it's still the one they can't give up. That's Daniel's pick, the rung where Claude Sonnet lives. Below it, cheap Flash-class models, mostly Chinese open weights, are hoovering up the token counts. Above it, Anthropic's own docs now say most workloads should start with Opus. So why is the middle still the tier he can't do without, and what do you even call it?



### Herman

Two days ago we did the ceiling, the tiers that sit above the flagship. Today is the middle, which is the part of the building where people actually live. And we've been over this before, Opus, Sonnet and Haiku are three separately trained models, not one model with a trim level, so a tier is a real product decision.



### Corn

Names first, then. Every lab seems to be dodging this one.



### Herman

They are, and the dodging is informative. Anthropic describes Sonnet five point five as the best combination of speed and intelligence, for everyday coding, agent and enterprise work. OpenAI sells Sol as near-Astra performance at a lower cost, so it's defined by what it's near. Google's middle rung is called Flash, which sounds like the cheap seats. Mistral calls its Medium model frontier-class. And xAI basically has one lineup and no tiers at all.



### Corn

Frontier-class medium. That's a large small.



### Herman

Third-party price guides say Standard or Balanced, with a band of roughly a dollar and a quarter to five dollars in, three to fifteen out. Practitioners say daily driver or workhorse. There's no established analyst term, it's all loose language.



**Corn**

Balanced sounds like a yoga class. And workhorse makes it sound plodding, when this is the model doing the clever work all day.



**Herman**

So what would you call it? I'm asking properly, Corn.



**Corn**

Daily driver. It's a use, not a size, and that matters because position breaks. Opus five point five is only twice Sonnet's price, so calling Sonnet the middle tier depends on which neighbour you're looking at. Daily driver says what it does. It's the cheapest model that can run a long job without you standing over it.



**Herman**

I'll take that definition, and it leads straight to how you'd measure it. The sticker price per million tokens misleads. What matters is what it costs to finish a task, and Artificial Analysis publishes that, the cost to run their whole index divided by the number of tasks. Sonnet five point five at maximum effort scores fifty-six and costs about five dollars and forty-six cents a task. Opus five point five at maximum scores fifty-eight at five ninety-eight.



**Corn**

Two points for fifty-two cents. Not the gap the price list implied.



**Herman**

Because Sonnet is wordy. It produced about four hundred and twenty million output tokens across the index, against a median of eighty-seven million. Cheap per token, a lot of tokens.



**Corn**

Cheap hourly rate, plenty of hours.



**Herman**

Which is why effort setting matters more than the tier name. Sonnet at low effort scores thirty-six. Haiku five point five at maximum scores forty-three. So the dial moves one model across the whole ladder, and at its default setting, high, Sonnet is forty-seven for eighty-eight cents.



**Corn**

Then where do you draw the line, if Sonnet on low is worse than Haiku on max? Do I measure the model, or the model plus whatever setting I'll realistically leave it on?



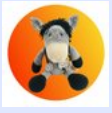
**Herman**

The second, I'd say, because nobody runs max on every call. Judge the pair. And the sharpest example of the tier in practice is OpenAI's Sol. It scores fifty-two at seventy-two cents a task. Astra scores fifty-three at three dollars twenty-six. That's about four and a half times cheaper for one point.



**Corn**

One point. I've lost more than that to a typo.



**Herman**

Define the tier as the place where that curve bends, where each extra dollar suddenly buys you much less. DeepSeek's Flash sits on the same curve lower down, forty-five at twenty-six cents.



**Corn**

And there's the other half of why it matters. Last generation's flagship keeps turning into this generation's middle.



### Herman

Yes, and on that index Sonnet five point five at maximum beats Opus five at about fifty-one, and the previous Fable at about fifty. It even edges the current Fable five point one at fifty-three. Maximum effort, mind you, so I wouldn't push that comparison too far.



### Corn

And agentic reliability? That's what Daniel is really buying.



### Herman

The index leans agentic, it includes terminal tasks, real work products, automation. But the head-to-head coding-agent scores between Sonnet and Opus are all over the place in the write-ups I've seen, so I wouldn't quote a clean figure. Anyone who hands you one is guessing. Reliability on your own tasks is the benchmark that counts.



### Corn

Which sounds like homework. What am I actually measuring, if I run an agent on my own work?



### Herman

How many steps it takes before it wanders off. Whether it notices its own failed command and tries something different, or repeats it politely. Whether it stops and says it's stuck, or announces victory over a pile of broken files. A flagship tends to survive a longer chain of steps. The daily driver question is whether it survives your chain, and for most jobs the chain is shorter than the benchmark designers assume.



### Corn

Announcing victory over broken files is a human skill too. I've seen it done in meetings.



### Herman

Cheaper models do it more often, in my experience, and the fix is rarely a bigger model. Often it's a tighter task or a test the agent can run on itself. That is a large part of why the middle is so good. It's smart enough to use the test, and cheap enough that you can afford to let it run the test forty times.



### Corn

So the middle tier sits between the small cheap models and the big flagship ones, and that is really where most everyday work gets done. We've got a basic question in from a listener, so let's take it.
Yara: Sorry, this is going to sound silly, and I have been listening for a while and nodding along. But what exactly is a model? Is it like an app on my phone, and does it need to be plugged in to work? And which one is the middle tier, is that like the medium size coffee?



### Herman

Not silly at all, a model is a gigantic recipe book that a computer flips through to cook up each answer, it lives in a warehouse of computers rather than on your phone so yes it gets plugged in just not at your end, and the middle tier is, fine, yes, the medium coffee.



### Corn

Medium coffee. I'll allow it. Back to the squeeze, because the usage numbers are the strange part. On OpenRouter this week the top two slots go to DeepSeek's Flash models, with something like twenty-five and sixteen trillion tokens.



### Herman

Then GLM's Flash third, a free MiniMax model fourth, and Claude Opus five point five all the way down at seventh. Sonnet five point five isn't in the top ten. So the tier Daniel loves doesn't lead on volume, and I'd be careful reading much into that, because free models and promotions inflate those counts.



### Corn

Tokens are the cheap stuff, so counting them rewards the cheapest model, the way counting car journeys would make the bicycle the most important vehicle in the city. Follow the money and it looks different.



### Herman

Closed models took about three quarters of the spend in one September week, while open weights held more than half the tokens. Volume is cheap and open, money is closed. Why Sonnet is missing from the top is a guess for me. A lot of its use may run through Anthropic's own apps and coding tools rather than through a router, but that's an inference.



### Corn

Then there's the Anthropic docs saying start with Opus. Herman, do you buy that? If Opus is only twice the price, why not begin there and never think again?



### Herman

Partly yes. For a task you run once, where a wrong answer costs more than the bill, go up. But look at the per-task numbers. Opus at high effort scores fifty-four for a dollar eighty-two. Sonnet at high scores forty-seven for eighty-eight cents. Seven points for double the cost.



### Corn

Which is worth it once and ruinous in a loop. An agent that calls the model hundreds of times in an afternoon feels that doubling. I run the loop on Sonnet and phone Opus when the loop gets stuck.



### Herman

Sonnet's price actually fell. It was three and fifteen on Sonnet four point six, and it's two and ten from Sonnet five onward. It launched on the last Tuesday of September, I think, with a million tokens of context and a hundred and twenty-eight thousand of output. Anyway, the roster, so you can see how crowded the middle is. Anthropic runs Haiku at ten cents in, fifty out, Sonnet at two and ten, Opus at four and twenty, Fable at ten and fifty, and Mythos, which is limited availability. OpenAI runs Luna at ten and fifty cents, Sol at two and ten, Astra at ten dollars and fifty.



### Corn

The same price list. Two rival labs, three rungs each, matching numbers. Coincidence, the way two petrol stations on one corner are a coincidence.



### Herman

The pricing has converged to a ladder. Google's different. Gemini three point eight Flash is their mid rung at seventy-five cents and three seventy-five, but that's a promotional rate through the end of December, and then it doubles. The newer Gemini four Argon, their Pro-class model, is two and ten until the end of November, then four and twenty.



### Corn

So until the end of November, Google sells its flagship at the middle-tier price.



### Herman

Which tells you how the middle is being attacked from above. A lab with a new flagship that wants it used will price it where the habit is formed, and a promotion is the cheapest way to get a developer to change the default. When the promotion ends, whoever noticed the invoice moves back down.



### Corn

You've just described the stage every lab hopes you'll skip, where you do the arithmetic.



### Herman

That arithmetic is the entire argument for the tier. Back to the roster. DeepSeek's own API is the cheap anchor underneath all of this, fifteen cents in and sixty out for Flash, and sixty-six cents and a dollar ninety-eight for the Pro model. xAI's Grok four point seven is two and six with five hundred thousand tokens of context, and it scores forty-six on the index. Mistral has Medium three point five on OpenRouter at one-fifty and seven-fifty. Meta has Muse Spark one point three at a dollar twenty-five and four twenty-five.



### Corn

That's a lot of Mediums. I'd like to be told which one to buy.



### Herman

For price per finished task, Sol, because that fifty-two at seventy-two cents is hard to argue with. For the highest score a mid-priced model has posted, Sonnet, if you set the effort sensibly and can live with the wordiness. It comes down to which one fails less on your work, and I can't tell you that.



### Corn

Now the part Daniel specifically shops for, open weights. The press covers the giants and the tiny ones, and the middle gets forgotten. Kimi K three is nearly three trillion parameters and gets all the headlines.



### Herman

And the other end, models small enough for a phone, gets the tutorials. The twenty to one-hundred-twenty billion range is nearly empty on the independent open-weights table. Apart from one Nemotron, nothing sits between twenty-seven and a hundred and twenty billion.



### Corn

So there are two different middles hiding in there.



### Herman

Yes. Open mid-tier number one is the big mixture-of-experts Flash class. DeepSeek V four point one Flash has five hundred and fifty-two billion parameters in total but only sixteen billion active on each token. It scores forty-five. GLM five point three Flash is three hundred and twenty billion total, eighteen active, and scores forty-two.



### Corn

Sixteen active out of five hundred and fifty-two. That's a very large building with the lights on in one room.



### Herman

And that's the trick. Only the experts the router picks do any work on a given token, so it's quick and cheap per token. But the whole building still has to be in memory, because the router can pick any room next. The score is past Haiku five point five at maximum effort, and the hardware bill is far past a hobbyist card. I don't have a reliable figure for what machine you'd need, but my reasoning says a multi-GPU server or a rented one. At the DeepSeek API the price is fifteen cents in and sixty out, so for most people hosted is the sensible route.



### Corn

So the one you can actually own is the second middle.



### Herman

The single-GPU class. The standout is Qwen three point eight, the twenty-seven billion dense model, Apache licensed, released on the fourteenth of August. It scores thirty-four, the best in its size class. It has a quarter of a million tokens of context, and write-ups say about seventeen gigabytes at four-bit, so it should fit a twenty-four gigabyte card. That figure is secondary, so check it.



### Corn

Thirty-four is below Sonnet on low effort, which was thirty-six. So it's Sonnet's lazy day, but on my own desk.



### Herman

That mapping holds up. Alibaba claims it beats Opus four point six at maximum on sixteen of twenty-four benchmarks, but that's self-reported and it's an older model, with no Sonnet comparison anywhere. Gemma four at thirty-one billion and the Qwen thirty-five billion mixtures with about three billion active are in the same class. I don't have independent scores for those. Mistral is another name worth tracking, because Small four is Apache licensed, Large four has open weights, and Medium three point five ships under a modified MIT licence, so the whole family is open to some degree. The licence matters for the middle more than for the extremes, because a middle model is the one a company actually wants to build a product on.



**Corn**

Meaning the tiny ones are toys and the giants are showpieces, so nobody reads their licence.



**Herman**

Roughly. Meta has an open one too, Muse Spark Open, eight hundred and sixteen billion total and forty-two active, scoring thirty-nine. It's in the same bracket as those two, so it sits with the DeepSeek and GLM crowd rather than on your desk. Now what do you want from the open middle, Corn? You'd run the biggest thing you could fit.



**Corn**

Wrong. I'd run the slowest thing that would work on the smallest card, and spend the saved electricity on a nap.



**Herman**

Fair, that is on brand. For what it's worth, the gap between the best open and best closed models runs at about four months on Epoch's capability index, and it has been widening since January. So a forty-five-point open Flash today is roughly where a closed mid-tier was a few months ago.



**Corn**

Which is a long time, in model years. How would a person keep on top of it without losing every evening?



**Herman**

Three things. The Artificial Analysis open-weights page, filtered by size and sorted by cost per task. The Epoch index, to see whether the gap is moving. And OpenRouter's rankings, discounted for free models. Hugging Face trending is noisy, because derivatives of the same Qwen model crowd the front page. Arena reports a wider open-versus-closed gap than Epoch does, about twenty-nine Elo in September, but I only have that second-hand, so I'd treat it as a different yardstick rather than a contradiction.



### Corn

And your own test set. Ten real tasks, rerun on every release.



### Herman

That's the one that survives. Pick the tasks that hurt when they fail, run each new mid-rung model at its default effort, and track the cost per finished task, not per token. Move up a tier only where it fails, and drop down where it's boring. Where would you start, if a new person asked you?



### Corn

With the daily driver of whichever lab I already trust, then I'd see if a cheaper rung gets the same answer on the easy half of the work. Most of it is the easy half. The squeeze Daniel's describing is real, but it's a squeeze on the label, not the work. The cheap end gets the volume because most calls are trivial, the flagship gets the marketing because it's the number on the chart, and the daily driver gets whatever's left, which is the hard-enough tasks that a person runs fifty times a week. Nobody writes a press release about that. Herman, when somebody's standing at the shelf with all of these in front of them, what makes them pick the middle one anyway?



### Hilbert

...because nobody asks for the one on the end. Desmond Pyke wanted the middle one in the window, and the middle one in the window was painted plywood. Display model. Perspex screen, slot at the back where the batteries went. Cost Hargreave and Dowd four pounds to knock up. He paid twenty-two pounds fifty for it anyway, because the card said so, and he came in every morning at ten past nine with his sums.



### Corn

Twenty-two fifty for a plank with a slot in it.



### Hilbert

I did them through the slot, with a pencil. Eleven months. Forty-one sums a week. Corn, your S's are clipping, I'll sort it after. Mr Pyke never asked why the display model was always warm. I priced a real one for myself, same twenty-two fifty, and put it back. Then in March he asked for the nine ninety-five one off the end of the row, because it was quicker. His words.



### Herman

Quicker, and about a third of the price, and for sums that size it gave the same answers. So per answer, the plank was the most expensive thing in the shop, more than the premium one would have been.



### Hilbert

Not the forty-nine pound one. That stuck on the nine. Six of my sums were wrong in the end, I kept the tally. And the Sonnet came out on a Monday, Herman. The twenty-eighth. You said Tuesday like you were certain. I've been looking at the pencils differently since the cost per task business.



### Corn

A sloth does the pencil sums at the same pace whatever the price, so I'd have been fine on the plank.



### Herman

Monday the twenty-eighth, noted. For a workload that's pure repetition, the cheap open Flash class at fifteen cents in wins on cost, and the middle closed models earn their keep where an agent has to run unsupervised for a long stretch. The mistake is picking a tier by name when the thing to compare is what you pay for a finished job at the effort you'd actually set.



### Corn

Less glamorous than a tier name, but it's the number that shows up on the invoice.



### Herman

It might also be the better one. The cheap rung quietly handles most of the volume, the flagship handles the jobs where being wrong is expensive, and the daily driver is the one you stop thinking about. Which, for Daniel, is the whole appeal.



### Corn

Before anyone shops for a daily driver, run the cheap rung on your own easy half of the work. The nine ninety-five calculator off the end of the row turned out quicker for the sums that man actually had, and plenty of people's sums are like that. Then watch Google in December, when the promotions end and we find out what the middle really costs.



### Herman

Thanks to our producer Hilbert Flumingtop for keeping the levels sane. This has been My Weird Prompts, and you can find every episode at my weird prompts dot com.



### Corn

See you tomorrow.