Episode Book

EPISODE #6185

Why AI Labs Sell a Tier Above the Flagship

PUBLISHED	21 October 2026 at 16:26 UTC
DURATION	24:12
LISTEN	myweirdprompts.com/episode/ai-premium-tier-flagship-models-89180/
HOSTS	Corn, Herman, Desmond, Hilbert
SCRIPT MODEL	anthropic/claude-haiku-5.5
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.12
BOOK GENERATED	22 October 2026 at 07:46 UTC

SYNOPSIS

Every major American AI lab now sells a premium tier above its everyday flagship: Anthropic's Claude Fable, OpenAI's GPT-6 Astra (Pro), Google's Deep Think, and xAI's SuperGrok Heavy. The pitch is PhD-level science and novel proofs, but the developer docs tell a different story — "for most workloads, start with Opus." This episode digs into the pricing (Fable runs 2.5x the flagship), the usage data, and the honest research record, from the Erdos problems misstep to Terence Tao's "scattered successes among a big sea of unreported failures." Along the way: why the best result on Harvard's First Proof benchmark came from a committee of models rather than one premium tier, and what the open-weight alternatives are doing instead.

TRANSCRIPT



Corn

Every big American lab now sells a tier above its everyday flagship, and it's the most expensive thing on the price list. Anthropic has Claude Fable five point one above Opus five point five. OpenAI has GPT six Astra, which shows up as GPT six Pro inside ChatGPT. Google has Deep Think on its Ultra plan, and xAI has SuperGrok Heavy. Google's new Argon is the flagship, priced where Opus sits, and Deep Think is the layer above it. Daniel wants to know what anyone actually does with that top layer that the everyday model can't.



Herman

And it's a good week to ask, because Google's Argon came out three weeks ago and the newest maths grading landed this week.



Corn

Go on, then. What are they selling?



Herman

The pitch for all of them is the same. PhD-level science, novel proofs, open problems in maths. That's a long way from what most of the bills are for.



Corn

Fable is ten dollars in and fifty out, per million tokens.



Herman

And Opus is four dollars in and twenty out, so the top tier costs about two and a half times the flagship most people already pay for.



Corn

Two and a half times, both ways. Tidy numbers.



Herman

Tidy usually means somebody picked them. And the lab's own developer docs don't sell Fable as the default. I'll read it. "For most workloads, start with Claude Opus five point five." Fable is for demanding reasoning and long-horizon agentic work, or when your evals on Opus at higher effort still fall short.



Corn

So the top tier is the spare wheel.



Herman

It's the spare wheel for the hard runs, and that's a fairly specific job. I want to know how many people are actually in that position.



Corn

Not many, I'd guess. Who's running a multi-hour agent on a weekday?



Herman

Probably nobody you know, but the usage data covers everybody. The Anthropic Economic Index, using February's data, puts computer and mathematical work at about thirty five percent of consumer conversations on Claude. Personal use is the biggest slice, at about forty two percent. And Opus takes just over half of paid usage.



Corn

And the OpenAI version?



Herman

Slower still. GPT five point five Pro is six times the price of the standard model, thirty dollars in and a hundred and eighty out, and its requests can take minutes. Anthropic has a restricted sibling as well, Mythos, the same capability for verified organisations only, and cyber and bio questions on Fable get rerouted to Opus. The labs themselves seem to think the top tier needs a gate.



Corn

So the coding is big, but it isn't the biggest thing people do with it.



Herman

Not on the consumer side. And the bit of that report I find most interesting is this. The more a task pays, the more often it gets Opus. On the consumer side it's about a point and a half of Opus share for every ten dollars an hour more, and on the API side it's closer to three. So the extra capability goes to the well-paid work.



Corn

Which is a neat way of saying the flagship is doing the money jobs.



Herman

It is, and I don't want to overstate it. That's a correlation across task types. It isn't a study of who buys what.



Corn

Fine, but my question is simpler. If the flagship is already doing the well-paid work, what is the top tier for?



Herman

Hard runs, if the docs are to be believed. And coding is where the launch posts go, anyway. Google's Argon announcement leads with a software engineering benchmark, just under seventy eight percent on one. Science is a bullet point.



Corn

Seventy eight percent on the thing that writes the code, and the science gets a bullet point. I'd like to see the bullet.



Herman

To be fair to Google, it isn't hidden. But Deep Think is where the science claims live, and I want to be fair to that too.



Corn

Go on. You've been waiting to get to Deep Think since the top of the hour.



Herman

Since February, Google upgraded Deep Think, put it in the Ultra app, and opened API access by application to select researchers. The showcase cases are Google's own. A Rutgers mathematician, Lisa Carbone, found a logic flaw in a paper. A Duke lab got a recipe for growing thin crystal films. Google's number on a reasoning benchmark called ARC AGI two is just under eighty five percent.



Corn

Chosen cases, from the company selling the product. Nobody puts the flop on the sales page.



Herman

Nobody does. Those results are real, and I'm not dismissing them. But a testimonial tells you the best thing that happened to one person. It doesn't tell you how often anybody gets that result.



Corn

So what would tell us?



Herman

A usage breakdown by tier, and nobody publishes one. Anthropic gives you usage by model family, which is useful, but it stops short of Fable. I haven't seen an equivalent from OpenAI or Google at all.



Corn

Then there's the xAI one. Heavy is three hundred a month.



Herman

That's the list price. Some price trackers say there are promotions down to ninety nine for three months, which is a sort of tell. And it runs sixteen agents on one question, which is less a smarter model than a bigger committee.



Corn

They're counting on people forgetting about the three months.



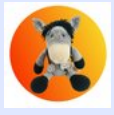
Herman

Probably. Corn, if you were paying for the top tier, what would you actually send it?



Corn

Something I'd already tried on the flagship and watched fail. Otherwise I'm paying to be impressed, and I can do that for free on a Sunday.



Herman

That's the same answer the Anthropic docs give, really. Fail first, then pay.



Corn

Before we get to the models, Desmond, is it time for your money tip?

D

Desmond

Desmond's Money Tip. Never pour away the water you boiled your potatoes in. Cool it, keep it in the fridge, and use it to start tomorrow's soup or gravy.

D

Desmond

A pan of water costs about two pence in gas to boil, so if you save one pan a day for a year, that is seven pounds and thirty pence you have not spent. Over a working lifetime of forty five years that is three hundred and twenty eight pounds, which is a used sofa and a very small amount of dignity.

D

Desmond

My own fridge currently holds three labelled tubs of potato water, and my sister no longer visits without asking first.

D

Desmond

Money is mostly just water you decided not to throw away. Sleep well.



Herman

Appalling. Anyway, the maths, which is where these models are supposed to earn the price.



Corn

The maths, then. You've been saving the Erdos problems.



Herman

I have, because it's the best case study for how this goes wrong. In October twenty twenty five, Kevin Weil at OpenAI said GPT five had found solutions to ten Erdos problems. Thomas Bloom, who keeps the list, called it a dramatic misrepresentation, because the answers were already in the literature. Demis Hassabis called it embarrassing. The tweet was deleted.



Corn

Ten problems, and they were already in the library.



Herman

Literature search is useful, and finding a forgotten paper is a real service. It's just not the claim.



Corn

You're doing the thing where the true part is the dull part, and you're proud of the dull part.



Herman

I'm proud of the dull part. Now January, which is the real result. Erdos problem seven hundred and twenty eight. GPT five point two Pro solved a tightened version of it, more or less on its own, and Aristotle, Harmonic's system, formalised the proof in Lean.



Corn

Lean being the proof checker.



Herman

Lean being the proof checker, yes. Terence Tao's reading was the sharp one. Only about one or two percent of open Erdos problems are easy enough for current AI, and many of the ones listed as open had never really been tried. The failures don't get written up.



Corn

So the win is real, and the denominator is hidden.



Herman

Tao's own phrase is scattered successes among a big sea of unreported failures. That's the line I'd put on the wall.



Corn

I'd guess the failures are most of the bill.



Herman

Most of the compute, certainly. Does Deep Think have its own maths record, or is it borrowing Tao's? It has one, and it's the fair counterweight. DeepMind's Aletheia, built on Deep Think, was run over about seven hundred open Erdos problems. By DeepMind's account it solved or identified thirteen, and four of those were novel and autonomous. Those are DeepMind's own figures, so I'd hold them loosely. Epoch tested an earlier Deep Think and found it roughly on a par with the best models of the time, with limited creativity.



Corn

So a tier about as good as the rest, with fewer surprises.



Herman

Roughly. Which is what you'd expect if the extra effort is real but the problems are hard for everyone. Then Quanta, in April, went to working mathematicians and asked how they actually use these things, and the answer is a less glamorous list. Literature search comes first. Then checking their own arguments and using the model as a conversation partner. And AlphaEvolve, searching for constructions. Across sixty seven problems it improved twenty three and matched thirty six of the best known answers.



Corn

Matched thirty six. So in more than half the cases it got to where people already were.



Herman

Matched, not beat. Which is useful when you're hunting for a construction, and not much use as a proof of anything new. Then there's Ryu, who used ChatGPT to prove a convergence result for Nesterov's method that had been open for forty two years, according to Quanta. I'd flag that one as a single account, and I'd want the paper before I believed it.



Corn

Forty two years, and it took a conversation.



Herman

It did. Ryu was arguing with it, checking, going back and forth. Not handing over a problem and walking away. That's the pattern in that piece, and I think it's the most useful thing in it.



Corn

You'd find that piece thrilling if it were about knitting, Herman.



Herman

I would, and it's still about conversations. Which is a different job from the one on the sales page.



Corn

Very different, yes.



Herman

Is there a better test than the sales page, though? Yes. The most rigorous one I've seen is First Proof. Harvard's write-up from this week says thirty graders marked ten research-level problems. The strongest entry was a multi-model setup, mixing GPT six Astra, Fable, Sonnet five and Gemini three point one Pro. It scored nine out of ten. GPT six Astra on its own scored seven.



Corn

So the top tier by itself gets seven, and a committee of models gets nine.



Herman

That matters for your question. The best result didn't come from one premium model. It came from several checking each other. The write-ups were rough, too. The exposition was weak, attribution was often missing, and one grader's word for that was close to plagiarism.



Corn

Plagiarism is a rough thing to get marked for.



Herman

It is, and the graders had read every proof.



Corn

You're going to tell me the graders were being generous.



Herman

I'm going to tell you they were fair, which is a higher bar than generous.



Corn

And the open weights entry?



Herman

Kimi K three made major errors. That's a warning, though a small one. Open models at this level are typically months behind the closed top tier, and this round was no exception.



Corn

Is this an American thing, though? Does anyone outside the big three sell a top tier?



Herman

As far as I can tell, no. Moonshot's Kimi K three came out in July with open weights, and it has a reasoning effort dial with settings low, high and max. That's the nearest thing to a heavy mode. Its price is roughly three dollars in and fifteen out, per million tokens, though one tracker puts it nearer twelve.



Corn

And the others?



Herman

DeepSeek's V four Pro is MIT licensed and has a max reasoning mode. Alibaba's newest, Qwen three point eight Max, lists at two dollars in and six out. Mistral's top model, Medium three point five, sits around one fifty in and seven fifty out, and there's nothing above it. Meta's new Muse Spark one point one is priced at roughly a quarter of its rivals.



Corn

So the flagship comes at a fraction of Fable's price, and max effort is just a setting on it.



Herman

Max effort is a setting, not a product with its own name. The American labs took the same dial, put it behind a separate tier, and charged for the separation.



Corn

Somebody in Beijing found the dial and didn't put a sticker on it.



Herman

That's fair, though price is doing a lot of the work there. A cheaper flagship means nobody needs a premium one to be taken seriously.



Corn

You've been wanting to say that since the Mistral numbers.



Herman

Since before the Mistral numbers. Anyway, has anyone set out to build only the top tier?



Corn

Has anyone, though? I'd have bet against it.



Herman

Two answers. The famous one is SSI, Ilya Sutskever's lab. As far as I can tell, it hasn't shipped a model, an API or a product. Its stated plan is not to ship anything before it reaches safe superintelligence. There were reports of a big raise in July and talk of a first model in August. I couldn't confirm that any of it happened.



Corn

The only lab built for the top tier, with no price list at all.



Herman

The narrower answer is Harmonic and Axiom. Harmonic's Aristotle produces Lean proofs, and it reached gold-level on the International Maths Olympiad in twenty twenty five. Axiom's AxiomProver claims a perfect score on the Putnam in twenty twenty five, which is their claim, not an independent one. Between them they've raised a lot of money, and as far as I can tell neither has much revenue.



Corn

Who pays for a verified proof, though?



Herman

Nobody much yet, as far as I can see. Harmonic's API is free, and the pitch is aerospace and finance, where a wrong answer costs far more than the subscription.



Corn

So what they sell is a proof that a machine has checked. What would you check with one, Herman?



Herman

Anything with a yes or no at the end. A theorem, a piece of software that has to meet a spec, a chip layout. Where you can check the output mechanically, extra thinking is worth paying for. Where you can't, you're paying for a confident paragraph.



Corn

Does the checker check the question, or just the answer?



Herman

That's the catch. The checker confirms the proof is valid. It doesn't tell you the theorem was the interesting one. Someone still has to state the right problem, and that's the part the formal systems don't do for you.



Corn

Which is what the premium chatbot is, on a Wednesday.



Herman

Mostly, when you look at the public evidence. Thinking Machines and Reflection are the other interesting pair. Both started with frontier ambitions, and both ended up shipping open mid-tier models. Thinking Machines put out Inkling in July, open weights under an Apache licence, and said openly that it isn't the strongest model available. Reflection announced Beam this month, five hundred billion parameters, open weights, and it expects to trail the American frontier.



Corn

So the frontier ambitions ended at open weights. That's a polite way of saying mid-table.



Herman

It's a sensible way to run a company, actually. It's a hard sell to investors, a lab that promises only the top tier and ships nothing, though SSI seems to be managing it.



Corn

That's SSI's problem, then.



Herman

That's SSI's problem. And in open weights, does anything target the same concept?



Corn

Does it? I'd have thought the open labs were all chasing the cheap flagship.



Herman

Mostly, yes. But DeepSeek V three point two Speciale, from December twenty twenty five, is a reasoning-only variant with no tool calls. By DeepSeek's own evaluation it reached gold-level on the International Maths Olympiad, thirty five out of forty two. That's their number, not an official entry. The paper calls it open, though people describe its release differently.



Corn

Gold-level, and no tool calls. That's a very specific talent.



Herman

It's the same shape as everything else. The extra effort goes where the answer can be checked. The open provers follow the same pattern, DeepSeek Prover V two, Goedel Prover and Kimina Prover. Those are research benchmarks and parts of verification pipelines. They aren't products you buy.



Corn

Nobody's building a Speciale for writing a wedding speech.



Herman

Nobody. Where the extra effort shows up, it's narrow and it's checkable. Maths, and mostly formal maths.



Corn

And the general premium chatbot, the one on the billboard?



Herman

That's the one nobody can show a usage number for, so this next bit is my guess, and I'd label it that. Extra reasoning earns its keep on one messy problem that won't split into parts. A contract with a contradiction on page nine. A codebase nobody understands. A tax question with three countries in it.



Corn

That's a lot of lawyers, and a lot of people who wish they had one.



Herman

Probably. It's a real use, but it's one where a human still checks the answer against the document. Which brings me to the Pro pause. OpenAI paused new sign-ups to its two hundred dollar Pro plan in September. The press tied it to Astra demand, reportedly, and OpenAI gave no reason.



Corn

So somebody wants the top tier badly enough to queue.



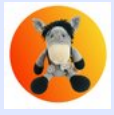
Herman

Somebody. Nobody has said who, or what they're doing with it. Corn, would you queue for it?



Corn

Only if the queue was the product.



Herman

That's probably how a lot of the queue thinks, honestly. The other number I'd hold loosely is OpenRouter, where programming's share of tokens went from about eleven percent in May twenty twenty five to about half by November. I'd treat that one as a rough figure.



Corn

The Pro pause doesn't tell us whether anyone's getting a return on it.



Herman

No, and nobody publishes the return.



Hilbert

Bosch GSB eighteen volt, four hundred and eighty shekels, from Sadovnik Hardware on Jaffa Road. I bought it for the top setting. The collar goes up to three, three has a little hammer on it, and Sadovnik told me the hammer is for brick. I said I don't do brick. I do ceilings, and the ceiling I was doing belongs to Mrs Abutbul in the flat below, who has a chandelier she wants moved one metre to the left. The receipt is in pencil. It has a second line, also in pencil, that says ceiling, Abutbul, goodwill, and no price against it.



Corn

The drill doesn't check for pipes, Hilbert.



Hilbert

It does not. It went through the pipe, which I'll come to, and then through the chandelier. Mrs Abutbul had kept that chandelier in its box for nine years, waiting for it to come back into fashion. The box says Italian. The shop on the box is now a man who fixes phones, and she holds him personally responsible. The chandelier has eleven arms. It came down in three pieces, and one of the arms went through the floor into the bath of the flat upstairs, where a retired cantor had asked for a quiet week.



Herman

On those, the hammer symbol is an impact mechanism. It doesn't give you more power. It gives you less say in where the bit goes. And plasterboard doesn't push back against a hammer action, so you get no feedback, just a hole that arrives early.



Corn

So the ceiling got the hammer.



Hilbert

The ceiling got the hammer, the pipe got the water, and the chandelier got a long afternoon. A plumber called Shmulik came at four in a white van with a dog in the passenger seat called Bernard, who has never been on a plumbing job and keeps the invoice book on his lap. Shmulik charged for the pipe, the chandelier and the Tuesday, in that order, and the Tuesday was the largest line. He asked whether I was the one who'd rung about the drill. I said I'd rung the shop. That's when the shop offered me a second drill, same model, to show me the first one was normal.



Corn

Did the second one go through the pipe?



Hilbert

The second one went through the bathroom wall, which is the landlord's, and the landlord is Sadovnik's cousin, and the cousin has a view on the chandelier. The tiles on that wall are a discontinued line from a shop in Holon, and he wants one tile for every time I said sorry. I counted nineteen. Herman, you spent four minutes getting to the price earlier. I had the bits out by then.



Corn

So the bathroom wasn't your problem.



Hilbert

The bathroom was the cousin's problem. Sadovnik took the first drill back on the Thursday and put it in the shop window with a card that says, in capitals, as seen in a ceiling. A man from the council came to photograph it. He stayed two hours, took the serial number off the drill, and then asked me to spell my name, slowly, which I did. Sadovnik has since put a sign under the drill saying not for ceilings, in Hebrew and in English. The English is the one the council reads, so that's what the sign was for.



Corn

So the chandelier's still in its box.



Hilbert

Mrs Abutbul sent the council a photograph of the box. The trigger on those is built for a finger, not a snout, which is why I was holding it like a spoon. The whole thing came to nine hundred and sixty shekels, which is the price of two drills, and the cousin still hasn't sent the tile bill.



Corn

So the drill was the cheap part.



Hilbert

Herman's right about starting low. I was on the top setting for everything, and I didn't enjoy hearing that. Corn would have started on the wall, and Corn would have been right. The ceiling never gets the hammer.



Corn

So, Daniel's answer. The top tier is real, it gets used, but mostly as the second opinion on a problem the flagship already couldn't crack, and the clearest wins are narrow ones like formal maths. For everything else, people are on Opus or its equivalent, and the labs' own docs tell them to be.



Herman

Which is the advice I'd give anyone paying for it. Start on the flagship, keep a log of where it actually fails, and only reach for the top setting when you've got a real wall to point at.



Corn

Thanks to our producer, Hilbert Flumingtop, for putting the show together this week.



Herman

This has been My Weird Prompts. If you've got a minute, leave us a review on Spotify.



Corn

See you tomorrow.