Episode Book

EPISODE #6382

Choosing AI Models on Cost and Performance

PUBLISHED	31 October 2026 at 15:45 UTC
DURATION	18:54
LISTEN	myweirdprompts.com/episode/model-cost-performance-tradeoffs-55735/
HOSTS	Corn, Herman, Dorothy, Neville
SCRIPT MODEL	anthropic/claude-sonnet-5.5
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
SOURCES	13 cited by the research (listed at the end)
BOOK GENERATED	1 November 2026 at 07:41 UTC

SYNOPSIS

Picking an AI model requires balancing benchmark scores against runtime costs, but popular formulas often produce completely different winners from the exact same data. A simple cost-to-performance ratio inevitably favors cheap models with unusable scores unless filtered by a quality floor. Exploring methods from Pareto frontiers and knee-finding to weighted sums and TOPSIS reveals why selecting the right evaluation formula is itself an unresolved multi-criteria decision problem.

TRANSCRIPT



Corn

Last time we split the models into price tiers. Now Daniel has added a benchmark score for each one, a number from zero to one on his own task, and he wants to know how to turn three everyday questions into arithmetic. Best performance. Best value for money, but only counting models above a minimum acceptable score. And best balance of performance and cost. Small table, a basic statistics calculator, no machine learning.



Herman

Herman Poppleberry here, and I'll say the good news first. These are old problems with decades-old methods. The bad news is that the methods disagree with each other on the same table.



Corn

Disagree how?



Herman

They crown different winners. Same data, different champion. We'll get there.



Corn

I'd have guessed the spreadsheet was the neutral party.



Herman

The spreadsheet is neutral. The formula you type into it isn't.



Corn

Fine. Where do we start?



Herman

With the easy one, because it shows what the other two lack. Best performance is a sort. Descending by score, take the top row. No weights, no normalising, nobody can argue with it. It's the only one of the three that doesn't depend on method.



Corn

Which is why nobody writes a paper about it.



Herman

Right. Best value is different. That's a ratio problem. Score divided by cost, or cost divided by score, dollars per point. One number per row, so it sorts cleanly. Best balance is the hard one. Two criteria pulling against each other, and no correct answer until you say how much you care about each.



Corn

And that one has an academic home.



Herman

Multi-criteria decision analysis, a branch of operations research. The standard description is that cost or price is usually one of the main criteria, and some measure of quality is another, easily in conflict with it. That's your table exactly.



Corn

So the ratio first.



Herman

The ratio first, because it's what the industry actually does. One of the public LLM scorecards blends input and output price at three to one, which is the convention for price-performance charts, then divides by the index score. Cost per point. And their headline is that the ranking changes completely once price enters.



Corn

Completely. Not shuffles. Completely.



Herman

Look at the spread. DeepSeek V4.1 Flash is about one cent per index point. Step 5 Preview, three cents. GLM-5.3, five. Gemini 4 Argon and GPT-6 Sol, eight. Claude Opus 5.5 and Kimi K3, fourteen. Claude Fable 5.1 and GPT-6 Astra, thirty-eight cents.



Corn

Thirty-eight times between the top and bottom of one table.



Herman

And nobody would call the Fable a thirty-eight-times-better model.



Corn

I'd call it a luxury purchase. I once paid extra for a leaf that was grown on a hill.



Herman

Did it do thirty-eight times the work?



Corn

It did the same amount of nothing, but with provenance.



Herman

Here's the flaw in the ratio. It has no floor. A model scoring point three at one cent beats a model scoring point nine five at five cents on points per dollar. Thirty points per cent against nineteen. The ratio loves the cheap one.



Corn

But you'd never ship the point-three model.



Herman

Unless point three is fine for your task, which is exactly the thing the ratio can't know. Think of a job like sorting incoming emails into "urgent" and "not urgent." A weak model might be perfectly adequate there, and paying for a strong one is waste. For drafting legal summaries, it's a disaster. The ratio sees no difference between those two jobs. So you add a floor. That's method two, and it's a filter, not a formula. Pick a threshold, say point eight. Throw out every row below it. Then rank the survivors by score over cost.



Corn

So the floor does the actual work.



Herman

All of it. And the floor is a judgement. Nothing in the table tells you it should be point eight rather than point six. A high floor means you're finding the cheapest model that clears the bar. A very low floor means you're finding the cheapest model, full stop.



Corn

Which means the same table gives different best values to different people.



Herman

Different tasks, yes. Take that scorecard. DeepSeek V4.1 Flash has the lowest cost per point, a cent, but it sits at forty-five on the index. Set any serious floor and it's gone. Muse Spark 1.3 is at forty-eight and about four cents a point, and that's the one they call best



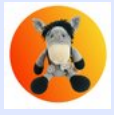
Herman (continued)

value at the frontier.



Corn

Same table, two defensible answers.



Herman

And both are right, given the floor. Which is the part people resist.



Corn

Why do they resist it? It seems obvious once you say it.



Herman

Because people want the table to settle the argument. A floor means you have to say out loud what "good enough" means, and that's uncomfortable. It feels like the number should tell you. There's also a geometric way to say the same thing. Plot score up the side, price along the bottom. Find your quality bar on the vertical axis, go straight right until you hit the line of best models, and whoever is there is the cheapest that clears it.



Corn

What line?



Herman

That's the next method. The ratio gives you a number, but it doesn't give you a shortlist. For that you want the Pareto frontier.



Corn

Spell it out. I know the name, I've never trusted it.



Herman

A model is dominated if some other model is both cheaper and at least as good. Throw out every dominated model. What's left is the frontier. Connect every model that nobody beats.



Corn

And the frontier picks the winner.



Herman

No, and this is the misconception. It eliminates losers. It hands you a shortlist, it doesn't rank. People keep expecting a crown.



Corn

So it's triage.



Herman

Honestly, yes, and the pediatrician in me likes that. It rules things out. It doesn't tell you what's wrong with the patient. Here's a worked case from a recent chart. GPT-6.1 Sol at max effort scores about fifty-two for roughly seventy-five cents a task. GPT-6 Astra at max scores about fifty-two and a half for about three thirty. Same score, give or take half a point, and four times further along the price axis.



Corn

So Astra is dominated.



Herman

Strictly speaking, half a point makes it technically undominated, but you'd never defend it. Which is its own wrinkle: dominance is exact, and your noise isn't. Run the same benchmark twice and half a point can easily move either way, so a strict rule is drawing a hard line through a blurry smudge. Then Grok 4.7 at xhigh, about forty-six and a half for three eighty. Five times Sol's price for a lower score. That one's plainly dominated.



Corn

Half a point on one benchmark run. I'd need a bigger number to leave the house.



Herman

Fair, and it's why people sometimes treat anything within a point or so as a tie and let price break it. That's a judgement too, but at least it's an honest one about the noise.



Corn

So the frontier is a rough sketch rather than a ruling.



Herman

A sketch that's still very useful. The practical advice from that same write-up is good. Take the two or three models on the line nearest your price, then run them on fifty to a hundred real examples from your own work.



Corn

Which is what we said about benchmarks in the ASR episode, that the score only matters if it measures your task.



Herman

Same lesson. Now, method four. The knee, or elbow. Where the curve bends from steep to flat. The point of diminishing returns. The standard algorithm is called Kneedle, from a 2011 paper by Satopaa and colleagues. Or you fit two straight lines and take where they cross.



Corn

Show me an actual bend.



Herman

The October frontier runs roughly seven cents a task for a score of thirty-eight. Thirteen cents for forty-six. Seventy-five cents for fifty-two. Six dollars for fifty-seven. Going from seventy-five cents to six dollars is about eight times the spend for roughly five more points.



Corn

That's the elbow, and I didn't need an algorithm to see it.



Herman

You didn't, no. Compare the first step: from seven cents to thirteen cents buys eight points. The last step costs eight times as much and buys five. But I'd warn you about it. Knee location depends on the scale of your axes. That criticism goes back to Worthing and Geffner in 1943. And on a log cost axis, the gap from ten cents to a dollar is drawn exactly as wide as the gap from a dollar to ten.



Corn

So I could move the knee by changing the graph paper.



Herman

Pretty much. And on a table of ten models, I'm not sure the knee means much at all. It may just be whichever point happens to sit at the bend. I haven't seen anyone test knee-finding on so few points.



Corn

Ten dots and a ruler. That's a vibe, not a method.



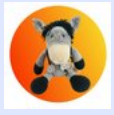
Herman

A vibe with a 2011 citation.



Corn

Well, before we get into the maths...



Herman

Oh, hold on. I think we have a call from your mother.



Corn

From Mum? Now?



Herman

Put her through. I'm delighted.

D

Dorothy

Hello dear, it's your mother. Dorothy's History of Toast.

D

Dorothy

In nineteen ninety-five a physicist called Robert Matthews published a paper explaining why toast really does tend to land butter side down. Mostly it's the height of the table, apparently.

D

Dorothy

And then the next year it won him an Ig Nobel Prize. So there you are, sweetheart. You drop your breakfast, and science explains it, and then it gets a prize, and the toast is still on the floor.

D

Dorothy

Anyway, I've got the kettle on. Speak soon.



Corn

Right. Weighted scores.



Herman

Method five. Normalise each column to zero to one. Score already is. Price becomes maximum price minus this price, over the range, so cheaper scores higher. Then U equals w times normalised score plus one minus w times normalised price. Sweep w from zero to one.



Corn

And the winner changes as I sweep.



Herman

At breakpoints, yes. Which is the honest way to answer best balance. Not one winner. A winner as a function of how much you care about cost. At w equals one you get the top scorer, which is just the sort from the start. At w equals zero you get the cheapest model. Everything in between is the interesting part.



Corn

That's a lot of effort to say it depends.



Herman

It depends, plus a chart showing where it flips. That's more useful than a shrug.



Corn

Fair. How would I read that chart in practice?



Herman

You'd look for long stretches where one model keeps winning as w moves. If a model wins from w of point three to point seven, it's a robust choice, because it survives a lot of disagreement about your priorities. If the winner flips every few hundredths, your answer is fragile and you should be honest about that.



Corn

Fair. But the normalisation worries me. Min and max come from whichever models happen to be in the table.



Herman

Add one absurdly expensive model and every other price gets squeezed toward one. They all start to look equally cheap, and the weights stop doing what you meant them to. That's why there's a rival, the weighted product model. It multiplies the criteria raised to their weights, so it needs no normalisation at all. It's more robust on mixed units, and it removes the bias normalisation can introduce.



Corn

And TOPSIS?



Herman

Hwang and Yoon, 1981. Imagine an ideal model that is best on everything and a worst one that's bad on everything. Pick the real model closest to the ideal and farthest from the worst.



Corn

So everything is a distance from something.



Herman

Roughly. Now, why the methods disagree. Triantaphyllou and Mann named it in 1989, the decision-making paradox. Different methods give different results when fed exactly the same problem and data.



Corn

Surely one of them is just wrong.



Herman

That's the intuition, and the paper's finding is nastier. When you use method X to decide which method is best, it says Y. Use Y, it says Z. And so on. AHP, WPM, ELECTRE and TOPSIS all show it.



Corn

So choosing the method is itself a multi-criteria problem.



Herman

Yes. Two concrete mechanisms in our table. First, ratios versus differences. A ratio rewards cheap models disproportionately; a weighted sum doesn't. Going from ten cents to five doubles the ratio, but it only moves a normalised price a little if the table has a six-dollar model in it. Second, scale dependence, which is the knee problem again.



Corn

So the method is the preference statement.



Herman

And that's not a defect. Picking ratio says you care about efficiency. Picking a floor says you care about a minimum. Picking a weighted sum says you'll trade one for the other at a fixed rate.



Corn

Does anyone argue the whole exercise is pointless?



Herman

Zellinger and Thomson, in an economic evaluation of LLMs. They say Pareto frontiers alone can't compare a cheap, error-prone model to a pricey accurate one. Their fix is to put everything in dollars: cost of a mistake, cost of delay, cost of abstaining.



Corn

And what falls out?



Herman

Reasoning models win on accuracy-cost tradeoff once a mistake costs more than a cent. Single large models beat cascades when a mistake costs as little as ten cents. And their conclusion is that you should usually just use the most powerful model, because deployment costs are dwarfed by the cost of errors.



Corn

That demolishes the whole value hunt.



Herman

Only if errors are expensive. If your task tolerates a wrong answer one time in five, hunt away. But that's a testable fact about your task. What would you do, Corn? Honestly. If a wrong answer cost you ten cents?



Corn

Pick the cheapest model that clears my floor and go for a nap.



Herman

That's actually the rational policy.



Corn

I contain multitudes.



Herman

Hilbert has been quiet, but I think our correspondent wants in. Neville, you're our correspondent on this.

N

Neville

Yes. Hello. I am, definitely, the correspondent. Well, intern. Uncle Hilbert's on his father's side, I'm on mine, that's the quills. He said you'd cover the Pareto thing and I should listen for the "pair of toes" part.



Herman

Pareto, Neville.

N

Neville

Oh, that's much better, actually. I looked up cheap models on the bus this morning. There was one for three pounds forty, on a leaflet, near Camden Road, and I picked it for my school project because it was cheapest. Did the knee thing. Sort of.



Corn

Did it clear your floor?

N

Neville

I didn't have a floor. Hedgehog spines are hollow and made of keratin, the same stuff as fingernails. Sorry. That just came up. Uncle Hilbert has a box for spare boxes, with a list of what's in it, and I used that list as my table. It had six columns. I got confused after four.



Herman

A table is a table, Neville. With six columns you'd pick the two that matter, score and price, plot one against the other, then cross out anything beaten on both.

N

Neville

That's the dominated bit! I mixed it up with the knee, Herman. The knee is where it bends.



Corn

Correct on both counts, in that order.

N

Neville

Brilliant. I'll tell Uncle Hilbert.



Herman

One thing his bus ride points at, and it's real. A price on a leaflet is a snapshot. The scorecard we mentioned has Opus 5.5 at fifty-four on high effort and fifty-eight at max. That's a four-point swing from the same weights, and a newer index version re-rated the whole field.



Corn

So an effort setting moves a dot further than a price cut does.



Herman

In that case, yes. The table you built last month may already be wrong.



Corn

Then the work happens before the spreadsheet. You decide what you care about, and then pick the method that says so.



Herman

And nobody has a principled way to set the floor. Every treatment I know of treats it as something you supply, not something you derive. That's a real gap.



Corn

Which leaves the question of how you'd derive it at all.



Herman

The frontier belongs to the set of models, not to any one of them, and the set keeps changing. So stop asking which model is smartest. Ask which one is on the line at your price, then test it on your own data.



Corn

Thanks to Hilbert Flumingtop for producing, and to Neville for sitting in.



Herman

For more along these lines, there's episode nine, Benchmarking Custom ASR Tools - Beyond The WER; episode five, Fine-Tuning ASR For Maximal Usability; and episode twelve, The AI Breakthrough. This has been My Weird Prompts. If you enjoyed it, leave us a review wherever you listen.



Corn

See you tomorrow.

SOURCES

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

1. **Zellinger & Thomson, Economic Evaluation of LLMs, 2025-07-04 [primary]**
   https://arxiv.org/abs/2507.03834v1
2. **Jung, Ricky & Chatni, 3D Optimization for AI Inference Scaling, rev. 2025-11-16 [primary]**
   https://arxiv.org/abs/2510.18905v3
3. **Giovannelli, Raimundo & Nunes Vicente, Pareto sensitivity, most-changing sub-fronts, and knee solutions, rev. 2026-09-26 [primary]**
   https://arxiv.org/abs/2501.16993v4
4. **Valdes Gonzalez, Cost-Aware Model Selection for Text Classification, 2026-02-06 [primary]**
   https://arxiv.org/abs/2602.06370v1
5. **Fowler et al., The Capability Frontier, 2026-06-25 [primary]**
   https://arxiv.org/abs/2606.26836v1
6. **Zhang et al., Beyond GPT-5: Performance-Efficiency Optimized Routing, rev. 2025-10-22 [primary]**
   https://arxiv.org/abs/2508.12631v2
7. **Wikipedia, last edited 2026-10-04**
   https://en.wikipedia.org/wiki/Multiple-criteria_decision_analysis
8. **Wikipedia, last edited 2026-09-14 (cites Triantaphyllou & Mann 1989; Triantaphyllou 2000)**
   https://en.wikipedia.org/wiki/Decision-making_paradox
9. **Wikipedia, last edited 2025-12-21 (cites Satopaa et al. 2011, Kneedle; Worthing & Geffner 1943)**
   https://en.wikipedia.org/wiki/Knee_of_a_curve
10. **Wikipedia, last edited 2025-10-23**
    https://en.wikipedia.org/wiki/Weighted_product_model
11. **Wikipedia, last edited 2026-05-27 (Hwang & Yoon 1981)**
    https://en.wikipedia.org/wiki/TOPSIS

12. **Aviral Bajpai, chart as of 2026-10-13**
https://www.aviralbajpai.com/blog/pareto-line-pick-ai-models.html

13. **iadecider, verified 2026-10-21 (table rows added 2026-10-26)**
https://iadecider.com/articles/llm-benchmark-scorecard-2026