Episode Book

EPISODE #6127

Haiku 5.5 at 10 Cents: What the Cheap Model Actually Does

PUBLISHED	19 October 2026 at 12:28 UTC
DURATION	22:25
LISTEN	myweirdprompts.com/episode/haiku-5-5-cheap-model-tested-36498/
HOSTS	Corn, Herman, Larry, Hilbert
SCRIPT MODEL	anthropic/claude-haiku-5.5
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.10
BOOK GENERATED	19 October 2026 at 17:43 UTC

SYNOPSIS

Haiku 5.5 launched on October 7 with a headline price of 10 cents per million input tokens — forty times cheaper than Opus 5.5 and a hundred times cheaper than Fable 5.1. But cheap tokens aren't the same as a cheap bill, and benchmark scores only tell you so much. We dig into the real numbers: token inflation, batch discounts, the 100K prompt-length pricing cliff, and what Anthropic's own comparison pages do and don't show. Then we test it against the job that matters most to us — writing this podcast. Haiku is already in our randomized trial pool, and one of its episodes aired. It scored well on vibe coding. It collapsed on agentic terminal tasks. It beat Sonnet on CyberBench. And when we asked it to write a script, it lectured. We walk through the eval data, the turn-length problem, and what it actually costs to run a show like this on a budget model.

TRANSCRIPT



Corn

Most of what gets said about Haiku goes the same way. It's the small one. It's what you reach for when you can't afford the proper model, and nobody puts the small one on the poster. So I'm going to take the other side, because Daniel's point is that Haiku is cheap, and he's right about that. Haiku five point five came out on the seventh of October, twelve days ago. Daniel wants to know what it's actually good at, what people hand to Opus that it could do just as well, and whether this show could afford to write its scripts with it, or whether the scripts would come out sounding like a leaflet.



Herman

I've been on the pricing page, Corn, and I'll say the surprising part plainly, because it's the whole point. Haiku five point five is ten cents per million input tokens and fifty cents per million output.



Corn

You've read that page twice.



Herman

Once, and then again to check it was real.



Corn

What's the proper model cost?



Herman

Opus five point five is four dollars in. And twenty dollars out. Sonnet is two dollars in and ten out.



Corn

Forty times the input price. And I'm the one who's meant to be slow.



Herman

The top of the range is Fable five point one, at ten dollars in and fifty out. So on input, Haiku is a hundred times cheaper than that. If you can wait, batch processing takes half off again.



Corn

Batch means waiting. For a sloth that's the same as free.



Herman

Anthropic also says it's around three quarters less to run than Haiku four point five, on average. Haiku four point five was a dollar in and five out. So they've cut the price tenfold, and then they told us about it.



Corn

Three quarters less, and they're still saying it. That's a confident discount.



Herman

There's a catch, and it's the one people miss. The same text comes out about thirty percent higher in tokens on this model than on the old one. So the bill falls by less than the sticker does. By my sum it's still several times cheaper for the same text. Just not a hundred times.



Corn

So the headline is the price of a token, and the bill is how many tokens you were going to use anyway.



Herman

Pretty much. Which is the whole argument for a cheap model. Volume.



Corn

Cheap, then. Now tell me what it's for, because the price doesn't tell me that.



Herman

Anthropic's own description is high-volume, cost-sensitive work. Classification. Extraction. Routing, which is deciding where a request goes. Summaries. Database queries. Live support chat, and browser use. They call it their fastest model to date, though they say it's less quick than their Opus models in Fast Mode.



Corn

Their fastest model, and it's slower than the expensive one on its fast setting. I can relate to that.



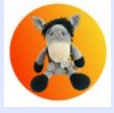
Herman

And they say it's built for subagent work. Their own copy says it pairs with Opus and Sonnet as a subagent on coding work. The big model plans the job. The small one goes and reads forty files and comes back with what's in them.



Corn

So it's the junior.



Herman

It's a job title, Corn. A subagent is the thing the lead hands the dull part to.



Corn

I've read forty files. I'd like someone to read them for me.



Herman

That's the whole idea.



Corn

Do you buy it, though? Is a subagent a different job, or the same job at a lower wage?



Herman

Partly. Reading forty files and reporting back is a job the small one does well. Deciding what the forty files mean for the plan is where I'd still want the big one in the room. Can I ask you something, Corn?



Corn

You're going to anyway.



Herman

When you pick a model for a job, what's your actual rule?



Corn

Whether I can stand to wait for the answer.



Herman

That's not a rule, that's a sloth.



Corn

It's a rule for a sloth. And it's how most people pick models, if they're being straight with themselves. How long they can stand the wait.



Herman

Then here's the evidence, and I want to pick the bits that make a point rather than read you the whole table. On the Vals AI vibe coding test, Haiku scores about ninety and Sonnet about ninety-two. Close enough that you'd barely see it in a day's work.



Corn

Two points. I can lose two points of anything before lunch.



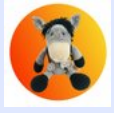
Herman

Terminal Bench is the opposite. That's the one where the model has to drive a real command line through a real task, start to finish. On Anthropic's own table, Haiku is around forty percent. Sonnet is around seventy. That's the agentic coding gap. And Anthropic says Sonnet and Opus remain the better choices for complex agentic coding. They're right to say it.



Corn

That's not close.



Herman

It isn't. Then there's CyberBench, on the Vals table, where Haiku is around seventy-five and Sonnet around sixty.



Corn

Then it beat the big one at something.



Herman

On one test, on one leaderboard. That's about how much I'd trust any single leaderboard, including the one I just read you. On Vals' overall index, Haiku comes nineteenth out of fifty.



Corn

Nineteenth out of fifty. I'd have had that framed.



Herman

And one more thing about the evidence. Anthropic's comparison page puts Haiku against Sonnet. It doesn't show Opus at all. So when someone tells you Haiku matches Opus, check which page they're reading, because the Opus numbers I've seen come from third-party write-ups, not from Anthropic.



Corn

So the Opus comparison lives in somebody's blog.



Herman

An aggregator's table, mostly. I haven't checked it at source, so I'm holding it loosely.



Corn

I'd guess the stuff people hand to Opus isn't the hard stuff. I'd guess it's the stuff they're scared of.



Herman

Mostly, yes. Summarising a long thread. Pulling the fields out of an invoice. Sorting support tickets. The first pass at a test file. Finding where a function gets called across a codebase. And Claude Code's built-in explorer runs on whatever the main model is, unless somebody configures it to use the small one.



Corn

People are paying the big rate to look for their keys.



Herman

Basically. Most people never change the setting, because it isn't the first thing you look for. The split, as I see it: where the job is well defined and you can check the answer, the small one holds up. Where the job is open ended and the wrong answer looks right, I'd still want the big one.



Corn

Open ended, and the wrong answer looks right. That's most of my afternoons.



Herman

Daniel's set his Claude Code up so research and script drafting go to Haiku subagents by default, with Sonnet only when a task warrants it. He didn't want to burn his Opus allowance on script writing. Each subagent can carry its own model setting, so the small one reads and the big one decides, in the same session. Claude Code's built-in guide agent runs on Haiku, which tells you the people who built it trust it with the questions where a wrong answer is cheap.



Corn

Or they're saving money.



Herman

Both. Nobody at a company is only one of those.



Corn

Sensible. Save the expensive one for the parts that need thinking. The big one's the manager, and the small one's the intern who reads everything first.



Herman

The intern who reads everything first is most of what a manager ever needed. What I can't find is how fast Opus drains a plan's allowance compared with the small one. The help page doesn't give a multiplier. A third-party guide claims one, and I couldn't verify it. So anyone who tells you Opus costs a specific multiple on a plan is working from a guess.



Corn

Everybody's working from a guess on the plans. The plans are a guess with a logo on it.

L

Larry

Imagine... a small thing. A quiet thing, sitting at the edge of the bag, waiting. It does not ask for very much. It has never asked for more than a few cents. Feel how light it is. Intelligence, reduced to the smallest weight a person can carry. It handles everything that needs handling, and it never once asks what you were thinking. Introducing the Crisp Clip. A subagent for your snacks. It does the dull part, so the big bag can rest. It reads the crisps. It has never once been wrong about the crisps, and it will never tell you the crisps are not the point. Less quickly than the premium clips, but so gentle. Available now, at a price so low we will not say it out loud, because the price is not the point.



Corn

Intelligence, reduced to the smallest weight a person can carry.



Herman

Please don't do that again. Anyway. The batch price is the bit I'd actually use. Batch is for the work that doesn't need an answer this minute.



Corn

So, could this show afford it?



Herman

Easily. Our script writer runs at about twenty-five thousand tokens in and five thousand out. On Haiku five point five that's about half a cent a script. On Sonnet it's about ten cents. On Opus with thinking, about twenty.



Corn

Half a cent. I've spent more than that on leaves, and they didn't work either.



Herman

The writer is around three percent of an episode. The whole thing comes to about sixteen cents, and most of that is the voices. If Sonnet wrote every script, an episode goes to about twenty-eight cents. Opus takes it to about forty-two.



Corn

Money's settled, then.



Herman

The money's settled. Any of them is affordable. The question was never the bill. There's one price trap worth knowing for a show like ours. Haiku five point five prices by prompt length. Once a request goes over a hundred thousand tokens, the input rate jumps to fifty cents a million, even if part of it is cached.



Corn

So the big prompt is the expensive bit.



Herman

Only if you stuff it. Our script prompt is nowhere near that size. A pile of past transcripts would be.



Corn

Our back catalogue is the one thing I'd like it to read least.



Herman

The window is a million tokens, so in principle it could hold every transcript we've ever made. The price is that hundred-thousand line. It's a window you're not allowed to open.



Corn

That's every window I've ever been near. Then it comes down to whether it's any good.



Herman

And here's the part I like. We have actual data on it, which is more than most people have when they pick a model. Daniel's own admission is that he picks on a hunch. So do I. I was sure, for about a week, that only the big model could hold a joke. I had no evidence for it. I just liked the name.



Corn

You liked the name.



Herman

It sounds like a mountain. Small things don't get taken seriously.



Corn

Data's better than a hunch, as long as the data is about the right thing.



Herman

First, Haiku is in our randomised trial pool. Most episodes go to DeepSeek. About one in four gets a weighted draw across the other candidates, and Haiku is one of them.



Corn

So it's already been writing us. Some of us.



Herman

Everything else in our pipeline already runs on a cheap model. Grounding, planning, metadata, intake. Nobody argues about those. The writer is the only text stage anybody argues about.



Corn

Because the writer is the only one that has to be funny.



Herman

Or the only one where somebody can hear the difference. One of its episodes aired. The typical turn was about forty words, which sounds fine. But Herman's longest turn ran to about a hundred and sixty words.



Corn

Herman's longest turn was always going to be a problem.



Herman

Thank you, Corn. And it read out a list of regions, one after another, in order.



Corn

It read the list. Out loud. Like a bus timetable.



Herman

That's the lecture problem. Long monologue turns, and lists recited back. It's the failure that kept coming up.



Corn

Which you do, Herman.



Herman

I do it when the prompt asks me to be thorough.



Corn

Then the prompt's to blame.



Herman

Possibly the prompt. I'm not sure about this part, but I'd guess a small writer fills a long length target with explanation, because explaining is the easy way to get words down. On the seventeenth of October we ran an eval. Haiku wrote between fourteen hundred and



Herman (continued)

fifty and three thousand words. The target was near thirty-eight hundred.



Corn

It's told to keep the turns short, and it still lectures.



Herman

It's told, and it doesn't always listen when the script gets long. A thing that's never tired also never gets bored of its own list.



Corn

That's the worst thing you've said about a machine, and I've heard you on the pricing. Short isn't the worst of it. On fun, it scored a one or a two.



Herman

Length is a count. Fun is a judgement, and the judge is the other problem. Haiku was also the checker in that harness, and it marked every plan point covered on the first sample.



Corn

A checker that passes everything has been told to be nice.



Herman

Which is a warning about letting a model grade itself. Nobody has a score for dialogue that's funny, as far as I can find. There's a score for code. There isn't one for the bit where you laugh.



Corn

Herman, do you think the funny comes from the model at all?



Herman

Honestly, I don't know. I think funny is mostly about what the writer is willing to get wrong on purpose. Haiku might be willing. I've no evidence that it is. On the other side, Opus five point five wrote twenty-two offline episodes, and Daniel found them more enjoyable.



Corn

The big one's funnier, then.



Herman

Or the same kind of writer with a different process around it. We don't know which yet, and that's the open question. I'd bet on the process a bit more than the model. I'd bet on it badly, though.



Corn

Either Haiku is a bad writer or we're bad at asking it. Those are the two bets.



Herman

There's a test we've never run. Haiku has adaptive thinking, steered by an effort setting, and the default is medium. Our test ran without it.



Corn

So we tested the small model with its eyes shut.



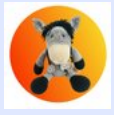
Herman

I was about to say that.



Corn

You were about to say it so I'd know you'd thought of it first.



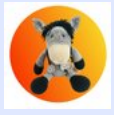
Herman

Turn the effort up and it thinks before it writes. For a script, that might be the difference between a plan and a monologue. Or it might just be a slower lecture.



Corn

A slower lecture. That's what I'd call a Tuesday.



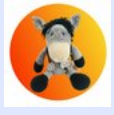
Herman

Corn, you've never had a Tuesday in your life.



Corn

I've had about four hundred of them, and I was asleep for most.



Herman

If I ran it, I'd do four scripts. Same brief, thinking off, thinking on, and the big model. Then we'd know whether the lecture is the model or the setting.



Corn

Four is a small number to bet a show on.



Herman

It's a fine number to test a model on. It's a small number to bet a show on, which is a different sentence. Corn, would you have got through the lecture part? Honestly.



Corn

I'd have napped through the second paragraph. That's my review.



Herman

And what would it take for you to believe a script was good? Not a score. You.



Corn

A line I didn't see coming, and didn't want to. Here's what I don't buy. You say the small one's fine where the job is well defined. A script is about the least well defined thing we make.



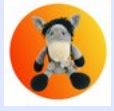
Herman

That's why I wouldn't give it the structure. But the format is well defined. Speaker labels, no digits, turns of a sensible length. A small model can follow a list of rules. It's worse at breaking them on purpose.



Corn

Breaking them on purpose is the job.



Herman

It's the bit people pay for, yes. So who gets the jokes?



Corn

Whoever's been awake longest. That's me. I'd still lose.



Herman

You'd lose to something that's never been tired. That's the unfair thing about them.



Corn

Herman, you'd put the Opus bill on everything if the allowance let you.



Herman

I'd put it on the bits where the thing has to be right. Not on the bits where it has to be fun.



Corn

Fun is the bit we're paid for.



Herman

I know. That's why I can't make up my mind about it, and I've tried.



Corn

Where would you trust it?



Herman

The boring pass. Did every speaker label survive? Is there a digit anywhere? Did a host quietly turn into a third host? That's a checklist. Checklists are the one thing a small model never gets bored of.



Corn

It'd be the proofreader, then.



Herman

A proofreader who reads everything. That's the forty-files job again, with a red pen.



Corn

Blind test, then. Same brief, four writers, and the listeners pick. Not us, either. Between us we've got a dozen hunches and no spreadsheet.



Herman

Fine. Blind, and nobody tells the listeners which one is which.



Corn

And what would make us switch the writer?



Herman

A blind test where the funny one wins, and the listeners can't tell which one is cheap.



Corn

The funny one, not the cheap one.



Herman

Those are the two columns. I'd like to see which one the listeners fill in.



Corn

Okay. The bit that matters is the bit we can't score.



Hilbert

So I said to Gershon, the goat is not on the quote. And he says, the goat is the quote. He's got the van parked outside my aunt's building, a white Peugeot Boxer with a bakery's name on the side, and it is not the bakery's van. The goat is called Dvora. She's grey and white, and she wears a harness Gershon made out of a seatbelt. The van does the street. The van does not do the stairs. It has one door that opens, and the door is the wrong door for a piano. The piano is my aunt's. It's a Bechstein. It's the wrong size for her flat and the right size for a boat. It has been stuck halfway up the stairwell since Monday, because Gershon was meant to get it up from the van in twenty minutes, which is a lie. It takes an hour and a half and two bottles of water. Now Dvora is taking it the rest of the way one leg at a time.



Herman

Lifting a piano by one leg is the problem. The frame carries the load through the legs. You're asking a cast iron frame to behave like a crane.



Hilbert

Gershon says the same thing. Then he lifts it by the leg anyway, because he doesn't read. Dvora does four hours on the stairs and gets it to the second floor, and she stops. She won't go above the second floor since the bathroom cupboard. The cupboard was not her fault. The cupboard was the cupboard's fault. Gershon's answer is the back stairs. They're narrower, and the piano doesn't fit them, and Gershon has measured that twice with a tape he keeps in his shirt. Dvora has views about the front stairs, because they have a rail, and she does not trust a rail. So the piano is on the landing, and the goat is on the landing, and the bakery's van is in the street, and nobody is going anywhere.



Corn

Four hours on the stairs. Dvora and I have the same pace.



Hilbert

She has her own invoice book now. She signs with a hoof, in green ink, on the bottom line. The bakery's man came round this morning to ask whether the van is still theirs. I tell him it's on an invoice now, under the goat. He stands by the van for a while, then he goes home. My cousin Maureen says the goat is a liability. Maureen runs a removals firm in Holon. She's been in the trade thirty years, and she is not to be trusted. She says a goat on a stairwell is a ramp with opinions. She says the only honest way to move a piano is a crane and a second piano to bang against the first. She says Gershon should be docked for the seatbelt, and Dvora should be docked for chewing the Peugeot's wiring, which is on the invoice, which is not on the quote. Maureen tells me a sofa is a boat in the middle of a move. She's sitting on it when she says it. She has been right about one staircase in her life, and it was hers, and she still talks about it. So weigh that. Gershon charges per staircase, not per hour. Nobody told me. It's on the invoice in small print, under the carrots.



Corn

Per staircase. Somebody finally priced it by the part that's hard.



Hilbert

Herman, you took the long way round to the price, and then the long way back. Gershon does that. He gives you a number, then he gives you the goat's number, and the goat's number is the one that happens.



Herman

Per staircase is the only number on that sheet that describes what actually happens.



Hilbert

Fine. Cheap for the stairs. I'll give you the stairs.



Herman

Cheap for the stairs is roughly where I'd put Haiku, if I'm straight about the forty-files job.



Corn

Then the scripts are the bit it isn't on the quote for. Per staircase, not per hour. That's the question I'd put to any model before you pick it. Per what?



Herman

And for the record, this script was written by Claude Haiku five point five, offline, in Daniel's Claude Code, from research notes that Haiku also gathered. So it's one data point, from the model under test. Judge it on whether it's funny, not on whether it was cheap.



Corn

Thanks to our producer, Hilbert Flumingtop, for keeping the desk running. This has been My Weird Prompts.



Herman

If you've got a minute, leave us a review. See you tomorrow.