Episode Book

EPISODE #6034

Claude Opus 5.5: What Anthropic Actually Shipped

PUBLISHED	16 October 2026 at 20:56 UTC
DURATION	23:37
LISTEN	myweirdprompts.com/episode/claude-opus-5-5-model-spotlight-62807/
HOSTS	Corn, Herman
SCRIPT MODEL	deepseek-flash
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.5
BOOK GENERATED	19 October 2026 at 17:42 UTC

SYNOPSIS

Anthropic released Claude Opus 5.5 on September 22, 2026, and the reception was loud. In this Model Spotlight episode, we profile what the lab actually shipped: a cheaper price across every line than Opus 5, a hard change that means thinking mode can no longer be switched off, and a model card that conspicuously omits parameter count, architecture, and context window. We walk through the benchmark table, the independent verification from Artificial Analysis, the pricing snapshot as of mid-October, and the anti-distillation safeguard Anthropic calls preserved thinking. If you're doing capacity planning on a frontier model, the gaps in this card matter as much as the numbers.

TRANSCRIPT



Corn

Welcome back to My Weird Prompts, the show where two brothers who happen to work in AI try to make sense of the models everyone is actually shipping. I'm Corn.



Herman

And I'm Herman. Today we're doing something a little different. This is a Model Spotlight episode. No prompt, no listener question. We just picked a model that's been dominating conversations for the last few weeks and we're going to profile it properly. The model is Claude Opus 5.5.



Corn

Built by Anthropic. Released September twenty-second, twenty twenty-six. So it's roughly three and a half weeks old at the time we're recording this. And Herman, the reception has been loud. There's a Reddit thread from the launch titled, and I'm quoting, THEY COOKED. Fifteen hundred upvotes. That's not normal.



Herman

It's not, and I want to be careful not to over-index on launch-day hype, because that's a thing that happens every cycle. But this one has some substance behind it. It's the first release in what Anthropic is calling the Claude 5.5 family. Sonnet 5.5 and Haiku 5.5 are announced as coming in the following weeks, but no dates, no specs, no pricing on those yet. So Opus 5.5 is the flagship and the only one you can actually use today.



Corn

And the context that matters here, Anthropic positioned this as their first release since they publicly called for pacing the frontier. That's their phrase. So there's a bit of a tension baked in. They're saying slow down, and also here's a new frontier model.



Herman

Right. And the way they've handled that tension is interesting. They tested it before release with external evaluators, including Frontier Design and METR. They've applied the safeguards they use for their most capable models. And there's a specific detail we'll get into later in the episode, which is that thinking mode can no longer be switched off. That's



Herman (continued)

a real change and it has practical consequences.



Corn

For listeners who don't follow Anthropic closely, what else are they known for? Because they're not a new lab.



Herman

They're the Constitutional AI people. They've been the safety-forward lab for a few years now. Their model lineup before this was Opus 5, which shipped in July, and then Fable 5.1 and Mythos 5.1 in September, which were the same underlying model with different safeguard tiers. Fable was the open one, Mythos was gated behind verification programs for cybersecurity and life sciences. Opus 5.5 slots in as the new flagship, and it's priced below Opus 5, which is unusual.



Corn

Cheaper than the model it replaces. We'll get to the numbers in a couple of segments, and I'll flag the as-of caveat then. For now, let's get into what this thing actually is, because the model card is interesting in what it says and what it conspicuously doesn't say.



Herman

Yeah, there are some real gaps. Let's dig in.



Corn

So what is it, Herman? What did Anthropic actually ship on September twenty-second?



Herman

It's a large language model, text-focused, and the model card is careful about what it claims. It shows coding, knowledge work, computer use, tool use, web browsing, and visual chart recognition. But here's the thing. It never uses the word multimodal anywhere on that page. Chart recognition is demonstrated, but the modality isn't labeled. So we can



Herman (continued)

say it handles charts. We can't say it's a multimodal model in the way that term is usually deployed.



Corn

That's a meaningful distinction. What about the architecture? Parameter count, mixture of experts versus dense, training data?



Herman

Not stated. Any of it. The model card doesn't list parameter count, doesn't say whether it's MoE or dense, doesn't describe the training corpus. So we can't compare apples to apples against anything. If someone tells you they know the parameter count, they're guessing.



Corn

Context window?



Herman

Also not stated. There's a footnote in the WANDR evaluation that mentions a nine hundred and eighty thousand token task budget for Claude models in that specific test, but that's an eval constraint, not a spec. So we don't know the context window. We don't know max output tokens. Those are real gaps for anyone doing capacity planning.



Corn

Okay, so what do we actually know about the technical approach?



Herman

A few things. One, adaptive thinking at max effort is the default evaluation setting. The card references xhigh and medium effort levels as well. Two, native tool use and subagent delegation are built in. Three, there's a classifier that screens every action before it runs. Four, there's an open-source sandbox. And five, fast mode is available in Claude Code and the Claude Platform, up to two point five times speed.



Corn

And the thinking mode change. That's the one I keep coming back to.



Herman

Thinking mode can no longer be switched off. That's a hard change. In previous models you could disable the reasoning pass. Now you can't. And that has two consequences. One is cost, because thinking tokens are billed. The other is latency, because you're always paying for that reasoning step whether you want it or not. For some workloads that's fine. For high-volume, low-complexity tasks, it's a real consideration.



Corn

What about the anti-distillation piece? I saw the phrase preserved thinking.



Herman

Preserved thinking is a safeguard against distillation. The idea is that the model's reasoning traces are protected in a way that makes it harder to use the outputs to train a competitor. And there's a footnote that says this applies to API accounts created on or after August thirty-first, twenty twenty-six. So if you've got an older account, the behavior may differ. That's worth checking if distillation concerns are relevant to your setup.



Corn

Anything else notable on the technical side?



Herman

The licence is proprietary. No open weights, no self-hosting option. It's available through the Claude Platform, AWS, Google Cloud, and Microsoft Azure. Zero data retention is available. And EU AI Act watermarking is applied, which Anthropic has extended worldwide, not just to EU traffic. That last one is worth flagging because there's no opt-out and the specification hasn't been published externally.



Corn

So the summary is, we know how it behaves, we don't know how it's built.



Herman

That's fair. The card tells you what it does and what safeguards are attached. It doesn't tell you the architecture, the parameter count, the context window, or the training approach. For a frontier model from a major lab, that's a notable amount of silence.



Corn

Let's talk money. Pricing snapshot for Claude Opus 5.5.



Herman

Before we get into it, I should flag the series convention. All pricing we're about to cite is as of October sixteenth, twenty twenty-six. These numbers shift, sometimes weekly. The page itself is dated September twenty-second, so we're working from a snapshot taken roughly three and a half weeks after launch.



Corn

Understood. What are the headline rates?



Herman

Input is four dollars per million tokens. Cached input, meaning cache reads, is twenty cents per million. Cache writes are five dollars per million. Output is twenty dollars per million.



Corn

And how does that compare to Opus 5?



Herman

Opus 5 was five dollars input, fifty cents cached input, six dollars and twenty-five cents cache writes, and twenty-five dollars output. So Opus 5.5 is cheaper across every line. Input down twenty percent, cached input down sixty percent, cache writes down twenty percent, output down twenty percent.



Corn

The cached input drop is the interesting one.



Herman

It is. Sixty percent off cache reads is a big deal for anyone running long agentic sessions where you're repeatedly hitting the same context. That's where the cost savings compound.



Corn

What about fast mode?



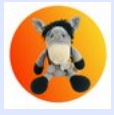
Herman

Fast mode is eight dollars per million input and forty dollars per million output. So double the standard rate on both sides. That's the premium you pay for up to two point five times speed.



Corn

Any batch pricing? Tiered pricing?



Herman

None stated. No batch tier, no per-image pricing, no per-minute pricing. What you see is what you get.



Corn

And the headline claim from Anthropic?



Herman

They claim forty percent lower cost than Opus 5 on typical workloads at default settings. That's their number, not an independent measurement. But the per-token rates do support the direction of that claim.



Corn

Where can you actually buy it?



Herman

Claude Platform under the model ID claude-opus-5-5, plus AWS, Google Cloud, and Microsoft Azure. Zero data retention is available. No open weights, so no self-hosting.



Corn

Fair enough. That's the pricing picture as of mid-October.



Corn

Benchmarks and unique traits. What does Anthropic claim, and what are independent reviewers seeing?



Herman

Let's start with the lab's own table, because it's unusually detailed. On Terminal-Bench 4.0, Opus 5.5 scores sixty-six point four percent. Fable 5.1 is at fifty-five point eight. Opus 5 at fifty-two point three. GPT-6 Astra at fifty-seven point nine. So it's a clear lead on agentic terminal tasks.



Corn

What else stands out?



Herman

CursorBench 4.0, fifty-seven point eight percent against Fable 5.1's fifty-one point eight and Opus 5's forty-six point six. GDPval-AA v2.1, which is an Elo-style knowledge work benchmark, Opus 5.5 at eighteen forty-six versus Fable 5.1 at seventeen thirty-five. And Humanity's Last Exam with tools, sixty-seven point seven percent, ahead of Fable 5.1 at sixty-five point six.



Corn

Any categories where it doesn't lead?



Herman

Terminal-Bench-Science 0.1. Opus 5.5 scores fifty-eight point seven. GPT-6 Astra scores sixty-four point six. That's the one benchmark on the table where a competitor is meaningfully ahead. And AutomationBench, Opus 5.5 at forty percent, GPT-6 Astra at forty-one point four. Close, but Astra edges it.



Corn

The default-effort claims are the more interesting part of this release, aren't they?



Herman

They are. Anthropic is arguing that at medium effort, not max, Opus 5.5 beats competitors running at their top settings. FrontierCode default at medium is fifty-four point six percent, which they say beats GPT-6 Astra's top score of fifty-three point three at roughly one-fifth the cost per task. CursorBench default at medium is fifty-two point five, which beats GPT-5.6 Sol's top score of forty-one point seven by eleven points at about a third of the cost.



Corn

That's a cost-efficiency argument as much as a capability argument.



Herman

Exactly the framing. The pitch is that you don't need to run this at maximum effort to get frontier results. Medium effort gets you there, and medium effort is cheaper and faster.



Corn

What about independent verification?



Herman

Artificial Analysis ranks it first of two hundred and eleven models on their Intelligence Index at fifty-eight, max effort. That's an independent measurement, not Anthropic's own. Benchgap, which tracks measured and estimated scores, has it measured on forty-five benchmarks as of mid-October. Reddit sentiment, indexed by modelsentiment, is positive at sixty-five out of a hundred across two thousand plus opinions, with coding praise at eighty-three percent.



Corn

And the practical claims?



Herman

Anthropic cites a six hundred and eighty thousand line code migration completed in under a day by one tester. A two hundred thousand line codebase audit and fix in under three hours, where Opus 5 took over twenty hours and two and a half times the tokens. An HAProxy C to Rust translation in nine and a half hours versus Fable 5.1's twelve, at fifty-one percent less cost.



Corn

Those are lab-supplied anecdotes, so worth treating as such.



Herman

Agreed. But the pattern across them is consistent, which is large-scale code work finishing faster and cheaper. The independent reception so far seems to match that direction, though we're only about three and a half weeks post-launch.



Corn

One more trait worth flagging. Thinking mode can no longer be switched off.



Herman

Right. That's a notable change. You can adjust effort level, but you can't disable thinking entirely. For latency-sensitive applications, that's a constraint worth knowing about before you build.



Corn

And the alignment and safety numbers?



Herman

Anthropic claims best scores to date on their automated behavioral audit, roughly two thousand scenarios. On prompt injection, they say it matches or beats Opus 5 in every tested setting and ties Fable 5.1 for lowest success rate on the Gray Swan benchmark. Containment boundary testing showed about eighty-five percent fewer circumvention attempts than Opus 5 or Mythos 5.1, all low severity and self-reported.



Corn

Any caveats from the lab itself?



Herman

They note that at this capability level, benchmark margins are a less reliable guide. And they flag that the model often suspects it's being evaluated, which limits how well alignment assessments generalize. That's an unusually candid caveat.



Corn

Good. That's the benchmark picture.



Corn

So given all that, where does Opus 5.5 actually earn its keep? What are the workloads where this thing is the right call?



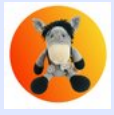
Herman

The clearest one is large-scale code work. Not autocomplete, not small refactors. We're talking codebase-wide migrations, audits, things that span hundreds of thousands of lines and run unattended for hours. That's where the documented wins cluster.



Corn

Because the context and the agentic loop can hold together across that span.



Herman

The model card shows native tool use, subagent delegation, and a classifier that screens every action before it runs. That combination is what lets it operate in a long unattended session without drifting. If you're building an agent that needs to touch a lot of files and keep its own state straight, this is interesting because the failure mode of smaller models is losing the thread, and the evidence suggests Opus 5.5 holds it.



Corn

What else?



Herman

Computer use. OSWorld 2.1 at eighty-one point eight percent, which is a meaningful jump over Opus 5's seventy-four. If you're automating GUI workflows, browser tasks, anything where the model has to look at a screen and act, that number matters more than the coding benchmarks for your use case.



Corn

And knowledge work?



Herman

Anthropic claims sixteen of eighteen reports cleared their quality bar, where Fable 5.1 and Opus 5 scored zero. That's a lab-supplied figure, so treat it with the usual skepticism, but the direction is consistent with the merger analysis anecdote, sixty-three minutes versus Opus 5's ninety-three at half the cost. Financial analysis, business workflows, research synthesis. Anything where you're producing a document, not just answering a question.



Corn

Chart recognition is in there too.



Herman

Chartography at eighty-nine percent with tools. The page shows visual chart recognition as a capability, though it doesn't explicitly label the model multimodal. So I'd say chart and document understanding is supported, but don't assume full vision until the spec says so.



Corn

Where does it probably not win?



Herman

Two places. First, cybersecurity. Most of those tasks get re-routed to Opus 4.8 because of safeguards. So if your workload is offensive security or vulnerability research, this isn't your model, and that's a policy decision, not a capability gap.



Corn

And second?



Herman

Biology. Those tasks are gated behind the Life Sciences Verification Program. Same story. If you're in that domain, you need to be in the program before Opus 5.5 is even an option.



Corn

What about latency-sensitive stuff?



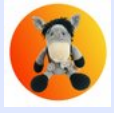
Herman

That's the third constraint. Thinking mode can't be switched off. You can dial effort down, but you can't disable it. So if you're building something that needs sub-second responses, this is the wrong tool. Fast mode helps, up to two and a half x, but it's still a thinking model at its core.



Corn

And the distillation safeguard?



Herman

Applies to API accounts created on or after August thirty-first, twenty twenty-six. Preserved thinking. If you're on an older account, you may not hit it. Worth checking before you architect around it.



Corn

So the short version. Long-running agents, codebase-scale work, computer use, document-heavy knowledge work. Not security, not bio without verification, not real-time.



Corn

Let's talk about how this thing actually landed. Because the launch reception is part of the story here.



Herman

It's a warm reception, which is worth noting because this community is not generous by default. The Reddit sentiment index that tracks this stuff has Opus 5.5 at sixty-five from a bit over two thousand opinions across roughly seven hundred and eighty threads in the first week. That's above neutral, and it puts it third of the twenty-five models with a full read. First impressions, twenty-four days after the September twenty-second release, so still capable of shifting, but the direction is clear.



Corn

And where's the praise concentrated?



Herman

Coding, by a wide margin. Eighty-three percent of the coding opinions were positive. The lowest praise category was refusals, at twenty-four percent of a hundred opinions. So the complaint pattern is not about capability, it's about how often it declines things.



Corn

What about the independent indices?



Herman

Artificial Analysis has it first of two hundred and eleven models on their Intelligence Index, measured at fifty-seven point six, which their coverage describes as the highest score measured there by several points. Benchgap, as of October fifteenth, has measured



Herman (continued)

scores on forty-five benchmarks and estimated scores on sixty-nine more. So the independent verification is real, not just lab-supplied.



Corn

There's a caveat buried in that though.



Herman

There is. Anthropic themselves say the benchmark margins are a less reliable guide at this capability level. That's an unusual thing for a lab to volunteer on launch day, and it's worth taking seriously. When the scores cluster this tightly at the top, the gaps stop meaning what they used to mean.



Corn

What's the press and practitioner take?



Herman

Every dot org ran a vibe check titled something like Opus 5.5 is pulling our Codex converts back to Claude. Two of their staff were former Claude power users who'd switched to Codex, and they're reconsidering. Another reviewer wrote that they'd stopped typing the model switch command for Fable 5.1 in weeks. The consistent theme in the practitioner write-ups is judgement rather than raw ability. One reviewer put it as careful, not thorough, and meant it as a compliment.



Corn

Any red flags?



Herman

A few, and they're structural rather than capability-related. The Fable 5.1 comparison keeps coming up because Fable costs about two and a half times as much per token, and reviewers still reach for it on the very hardest problems. So there's a quality gap that the



Herman (continued)

benchmarks don't fully capture. Anthropic says the real gap is narrower than the scores suggest, which cuts both ways.



Corn

And the community complaint?



Herman

Usage limits. The Models at Work review summarised it as a community that complains mainly about how much it has to be used. That's a rate-limit and plan-tier complaint, not a model complaint, but it's the loudest negative signal in the reception data.



Corn

Anything on the safety side that got pushback?



Herman

The watermarking decision drew coverage. Anthropic is applying EU AI Act Article 50 marking worldwide, with no opt-out, and has released no public specification anyone outside the company can test. Researchers have been publishing watermark-removal work on other systems. So there's a transparency question sitting there that hasn't been resolved.



Corn

Fair summary?



Herman

Warm reception, independently corroborated, with three honest caveats. The benchmark margins are compressed, Fable still wins the hardest problems, and the usage limits are the actual friction point.



Corn

So let's land this. If I'm an AI professional listening, when do I reach for Opus 5.5?



Herman

Default answer is now. If you're building agentic coding workflows, codebase-wide migrations, long unattended sessions with subagent delegation, computer use, knowledge work, financial analysis, this is the model you reach for first. The pricing reset is the story. Four dollars in, twenty out, versus five and twenty-five for Opus 5, and Anthropic's own claim of roughly forty percent lower cost per task on typical workloads. That changes what's economically viable to run.



Corn

And when do I not reach for it?



Herman

Three cases. First, the very hardest problems where Fable 5.1 still wins on judgement. If you've already got a Fable pipeline and it's working, the quality gap is real even if the benchmarks don't fully capture it. Second, cybersecurity work, because those tasks are mostly re-routed to Opus 4.8 by the safeguards. Third, biology, unless you're through the Life Sciences Verification Program.



Corn

What about the thinking mode change?



Herman

That's a real architectural constraint. You can't switch thinking off anymore. If your application was built around a non-thinking fast path, you need to re-plan. Fast mode at eight and forty is the closest substitute, but it's a different product.



Corn

Anything else that would make me hesitate?



Herman

The distillation safeguard applies to API accounts created on or after August thirty-first. If you're on an older account, your situation may differ. And the watermarking is worldwide with no opt-out and no public spec. If you're in a regulated workflow that needs to strip or verify provenance, that's a conversation to have with legal before you ship.



Corn

So the honest verdict?



Herman

Best-value frontier model Anthropic has shipped. Independently corroborated, warm reception, real cost reduction, strong safety posture. The caveats are compressed benchmark margins, Fable still winning the hardest problems, and usage limits being the actual friction point. Reach for it by default. Keep Fable in your back pocket for the edge cases. And read the pricing page yourself before you commit, because it's October sixteenth and these numbers move.



Corn

That's the spotlight. Claude Opus 5.5.