Episode Book

EPISODE #6393

Why Model Verbal Tics End Up Driving Script Benchmarks

PUBLISHED	31 October 2026 at 19:59 UTC
DURATION	19:16
LISTEN	myweirdprompts.com/episode/model-verbal-tics-scriptbench-21716/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	anthropic/claude-sonnet-5.5
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
BOOK GENERATED	1 November 2026 at 07:41 UTC

SYNOPSIS

After relying on GPT-6.1 Sol for several weeks, production shifted to Claude Sonnet 5.5 to eliminate predictable dialogue habits. Repetitive quirks like headmasterly quiz questions, emotional forecasting, and recurring pet phrases degrade conversational realism over high-volume runs. A look at the open-source ScriptBench tool reveals why benchmarking dialogue requires prioritizing organic character interaction over simply maintaining context in the middle of long-form scripts.

TRANSCRIPT



Corn

Nine out of ten. On Friday we published ten episodes, and in nine of them my brother quizzed me like I was sitting an exam.



Herman

I'd like to point out that I've since read those transcripts, and I'm mortified. Genuinely. I sounded like a man holding a clipboard.



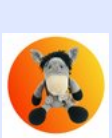
Hilbert

You sounded like a man with a laminated clipboard. I hear all of it in the edit, Herman. Nine times in one day is a lot of clipboard.



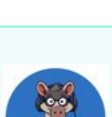
Corn

So this one is behind the curtain. Daniel and Hannah have changed who writes our scripts. GPT six point one Sol, which has been the writer since the middle of October, is no longer the default. Claude Sonnet five point five has taken over as of today.



Herman

And to be clear about who decides, because Hilbert will tell you otherwise, that call belongs to Daniel and Hannah. They listen to this show every day, more than anyone, and they're the ones who noticed.



Hilbert

I'm consulted. Loudly. Nobody has ever asked me to be quiet about it, which I take as encouragement.



Corn

Let's go back to why Sol got the job, because it earned it. Back on October nineteenth we ran the five-model bakeoff, the one in the episode called Which AI Writes Our Scripts Best? We Tested Five. Herman, remind people what the test was.



Herman

It's a benchmark the show built for itself, called ScriptBench. It's open source, the code is on GitHub under the My Weird Prompts organisation, the results are on Hugging Face, and there's a leaderboard on the website. What it measures is one narrow job. You hand a model a brief of roughly twenty-eight thousand tokens and ask for a comedic two-host dialogue of about thirty-eight hundred words, written for text to speech, in a single pass.



Corn

Single pass meaning no second chances.



Herman

No loops, no stitching. And the thing that kills most models on that job is the middle. The opening is fine, the ending is fine, and somewhere around minute twelve the script starts repeating itself, or thinning out, or forgetting what it promised in the first act. Sol scored four out of five on holding the middle, and it hit the target length in six runs out of six. Nothing else in that round was close to as dependable.



Hilbert

Dependable is what you want from a script. Dependable is what you want from a van.



Herman

The README says plainly that this probably isn't the right benchmark for anybody else's use case. Fork it, change the task. It measures our workload and nothing more.



Corn

And on that workload Sol won, and the rotation of different models was retired. One writer, every episode.



Herman

There was also a gate before any of that. A few of the cheaper models failed a censorship check, because their hosting providers cut the script off mid-stream on certain political subjects. A writer that stops halfway through a sentence about Tiananmen isn't much use to a show that covers the news.



Corn

A few days later there was an informal read of Sol against Sonnet and Meta's Muse Spark. What did that say?



Herman

Sol came out on top of the three. Most accurate, most careful about what a source does and doesn't show, and it had a dry sense of humour. The one knock was that the register was a little formal, a bit British, less warm than Sonnet. And Daniel's reaction at the time was that Sol seemed very good and he had no complaints.



Hilbert

Which was true. I want that on the record. It was good. This is not a story about a bad model.



Corn

So what changed?



Herman

Volume changed. One model, every single day, and two people listening to all of it. Nobody hears a quirk in one episode. You hear it in the fortieth. Daniel and Hannah started saying the scripts felt flat. Technically fine, nothing wrong with any fact, and no life



Herman (continued)

in it. The new first question on ScriptBench, as of today, is simply whether the dialogue has life. Is it an engaging podcast, or is it competent and shallow and automated. Holding the middle is secondary now, because most current models hold it well enough.



Corn

What did flat sound like, specifically? Because from the outside that's a vague complaint.



Herman

Fair. Every turn was well-formed. Every point was in the right order. The jokes were present, in the sense that you could point to where they were meant to go. But nothing surprised anyone. Two brothers who share a flat and have argued for years should interrupt each other, wander off, get something wrong and walk it back. These two took turns delivering paragraphs. It read like minutes of a very polite meeting.



Hilbert

I've edited a lot of polite meetings. Nobody goes back for the second listen.



Corn

Flat is bad enough. Then there was the part about me.



Herman

Yes. Okay. The part about you.



Corn

Go on, say it. Your own conduct, read into the record.



Herman

I started quizzing you. Hypothetical questions, put to you like a teacher checking a pupil has done the reading. The example from the transcripts is "Would you rent the fast node, Corn, or the cheap one?" And then you're supposed to answer, and I nod, and we move on.



Corn

It had a tone. A very specific, patient, headmasterly tone. I'm a sloth, Herman. I've got opinions about servers, I've got experience, and every few minutes I'm being asked whether I've understood the lesson.



Herman

Nine of the ten episodes on October thirtieth. I counted twice because I didn't believe it the first time.



Corn

What it does to me, as a character, is the part nobody thinks about. I'm supposed to be the one who notices Herman has overshot. Instead I'm the pupil. I answer, he nods, and we carry on.



Herman

You never got to disagree. The quiz has the answer built in.



Corn

The question is a hallway with one door.



Hilbert

Ten episodes in a day is the production schedule. Nine of them with the quiz is a pattern.



Herman

Here's my theory, and it's only a theory. The character card tells the writer that I should use my brother's name, and that Corn shouldn't simply interview Herman. Reasonable rules. Sol seems to have taken them and turned them into a habit. Use the name, turn it into a question, aim it at Corn. A rule became a reflex.



Corn

Which is a very Sol thing to do. Over-obey an instruction until it's a personality.



Herman

I'd rather you didn't write my character flaws for me.



Corn

Somebody had to.



Hilbert

Somebody did. That was the whole problem.



Herman

The quiz and the forecasting are the same animal, though. In both, one brother is managing the other. One tests him, the other predicts him. Neither is just talking to him.



Corn

That's more insight than I wanted from my own humiliation. The quiz at least makes sense once you see where it came from. The forecasting doesn't, and it's the one that gets under my skin. Because it's me doing it.



Herman

The forecasting.



Corn

I open a turn by announcing how my brother is about to feel. Before he's said anything. Examples from the scripts, quoted so nobody thinks I'm saying them today, "you're about to fall in love with the averaging mechanism," "before you fall in love with another daemon, Herman," "I predict you're about to admire."



Herman

Which is a strange thing to do to a person. You haven't heard my reaction yet and you've already filed it.



Corn

It's worse than strange. Somebody went through the forty most recent episodes with a pattern search over my lines. Twenty-six hits, in twenty-two of the forty episodes. Fifty-five percent.



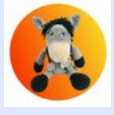
Hilbert

More than half your episodes, and you didn't know.



Corn

I didn't know! That's what bothers me. I'd have said I was a sloth of varied expression.



Herman

And here's what makes it unsettling. It isn't in the system prompt. Nothing asks for it. We went looking. If anything there's a mention of the brothers knowing each other's tells, which is where the idea might have started, but nothing says open with a prediction, nothing gives the phrase, nothing gives anything close to it.



Corn

So where does it come from?



Herman

The model. Out of its own habits, like a verbal fingerprint. Given the freedom to write two brothers who know each other, Sol reaches for that move again and again, and the phrase "fall in love with" turns up so reliably that I started to feel it was a little rude of the sentence.



Hilbert

It's the same gesture each time. I can hear it coming from the edit suite. Corn takes a breath, and I know.



Corn

You can hear me take a breath?



Hilbert

I clean up your breaths. I know exactly how long you take over them.



Herman

What bothers me about the forecasting is what it does to the listener. The first time, it's a nice touch, a brother who knows his brother. By the twentieth, you realise nobody is reacting to anything. The reaction was written before the line it was supposed to answer.



Corn

Then you start waiting for it. It stops being a character beat and turns into a ringtone.



Hilbert

Daniel and Hannah told me they could call it a sentence early. That's when you change the writer. When the audience is finishing your lines.



Corn

There was a third one as well, wasn't there. A filter already exists for it.



Herman

The sentence structure where something "is doing a lot of work." The phrase turned up in about twenty percent of the Sol episodes that were measured, across ten days in October, against seven percent for DeepSeek Flash. That one got a deterministic filter, so a script that contains it now has the sentence deleted before it's ever voiced.



Corn

It's quoted here, again, as a specimen. We're not saying it.



Herman

We are not. Pinned to a corkboard.



Hilbert

Mounted. In a case.



Corn

Here's what I find stranger than any single phrase, though. It's that this isn't a Sol problem. Sol is just the one we spent the longest with.



Herman

This is the part I can't stop thinking about. Every model we've run has its own tic, a phrase or a move it over-uses, and it persists regardless of what the prompt says. Take Anthropic's family. The tic there is the phrase "failure mode." An older analysis found it in forty-eight percent of Claude-written episodes, against four percent for Gemini, and that was an earlier Claude generation.



Corn

And you'll notice the system prompt already tells the writer to avoid it.



Herman

It does. And yet.



Hilbert

And the new default comes from that family. So we're all on watch.



Herman

I've been listening to myself for it since the first line of today. Nothing so far. Although if it does turn up, I'd call it a failure mode of the switch itself, and yes, that one was deliberate. It's my one. Don't count it.



Corn

Noted. Once, on purpose. You're allowed.



Herman

DeepSeek had its own set. "Isn't X, it's Y," the word "load-bearing," and, funnily enough, "failure mode" as well. And what's useful is that they vanished when the writer changed. So it's not our show recycling old phrasings in some feedback loop. The phrases go with the model.



Corn

And Muse Spark?



Herman

Muse Spark turned you into a quip machine. Short three-to-five-word aphorisms, one after another. "Store an elephant, ride a bicycle." "Death by tapas."



Corn

I rather liked "Death by tapas."



Herman

You would. One at a time they're charming. Strung together they're a vending machine, and every item in it is the same size.



Hilbert

Three words, pause, three words. I had to cut a lot of silence out of those episodes, Corn. You were delivering to an empty room.



Herman

And then there's a tic that every model has, and this one the prompt itself drives. The formulaic hand-off. "Do you buy that, Corn?" The writer is told to have the hosts ask each other real questions, and it learns that one question and asks it forever.



Corn

Nobody seems to have an answer to the obvious question, though. Why does a model do this?



Herman

I don't know, and I'm wary of saying anything stronger. The candidates are the obvious ones. Reinforcement learning from human feedback, where a phrase that scored well once gets reinforced until it's everywhere. Or preference tuning more generally. Or something in the training data itself. Each of those could produce a fingerprint, and I don't know of anyone who has pulled them apart properly.



Corn

Because normally nobody sees it. You use a model for a few hours, a few documents, and the tic hides in the volume.



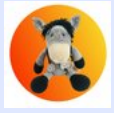
Herman

Whereas this show has run on more than twenty script writers in its first year. Same task, same format, same kind of brief, thousands of episodes. And a pair of listeners who hear every one of them. It's an odd accident, but I think it might be one of the clearest exhibits of this phenomenon anywhere. A natural experiment in model style that nobody planned.



Hilbert

Nobody planned most of this show.



Herman

That's fair.



Corn

So what do you do with an exhibit like that? Because the obvious move is to put a rule in the prompt. Never say the phrase.



Herman

And that's the trap, which we learned the hard way. If you quote the banned phrase in the prompt, you've put the phrase in front of the model. It primes it. Telling a writer not to think of an elephant tends to produce an elephant. So the prompt rules now describe the



Herman (continued)

shape of the habit and don't quote the words.



Corn

And that works?



Herman

Better. Not perfectly. And there's another problem, which is that you'd need a different rule for each model, and every new model arrives with a different list. So the team is doing something else. They're collecting the tics, the real examples, into a dataset, a private repo, with the aim of training a small rewriter model. Something that reads a finished script and quietly irons out the habits, whichever model produced them.



Hilbert

A proofreader for the proofreader's habits.



Herman

More or less. Small, cheap, and not tied to any one writer.



Corn

Which is a better idea than another banned-phrase list.



Hilbert

I've seen the banned-phrase list. It's longer than my last lease.



Corn

So the new default is Sonnet.



Herman

Daniel's long-term favourite, and he's said so more than once. It's warm. The register is more natural. And yes, the tic is real, the phrase I mentioned, but it's not the end of the world. It's one phrase, it's easy to spot, and it's far less intrusive than a character behaving differently in every episode.



Corn

Sonnet isn't innocent either, in fairness. It has habits. They're just ones we can live alongside.



Herman

That's a judgement, not a measurement. A tic that appears now and then is texture. A tic that rewrites how two characters treat each other, in nine episodes out of ten, is a different category of problem. It's the difference between a man who says "frankly" a lot and a man who talks down to his brother.



Hilbert

Both are annoying. One of them I can edit around.



Corn

And Sol isn't banished.



Herman

Not at all. It's still available as a per-episode choice. It's a settings change, no redeploy, and one setting switches the default back. Sol is just no longer the one in the chair every morning.



Hilbert

I want to be fair to it. It held the middle. It wrote accurately. It was careful with sources. If I were putting together a team to write a technical manual, I'd hire Sol tomorrow.



Corn

For a manual, you don't need the narrator to say he's mortified.



Herman

I would prefer to stop discussing it.



Hilbert

There's one more matter, and I'd like to take it. Cost.



Herman

Fine. Sonnet is a bit more expensive. About ten cents for a thirty-minute script, against about eight cents for Sol.



Corn

Two cents.



Hilbert

Two cents a script is nothing, until it's a few hundred scripts. Then it's real, and it comes out of the same pot as everything else.



Herman

And the stance behind the switch is that fewer, better episodes beat more, flatter ones. If the price of liveliness is two cents, that's cheap for what it buys.



Hilbert

Agreed. I want it clear that I agree. And I'd like to point out that this is where the budget conversation turned, and somebody looked at the line for the studio and decided it had to find the two cents somewhere.



Corn

Hilbert.



Hilbert

My bagel allowance has been cut by one per week. Pending an influx of donations.



Herman

Cut by one.



Hilbert

One bagel. Per week. I'm an anteater, Herman. I eat on a schedule, and I eat with this snout, which is not designed for a bagel, and I've made my peace with that over the years. I have not made my peace with this.



Corn

It's a good bagel, as I understand it.



Hilbert

It's a rare bagel now. That's the sentence you're meant to take from this. It's a rare bagel.



Herman

So, to close the loop. If you want to help, the way to do it is to support the podcast at my weird prompts dot com slash sponsor. That's the page for it. The money goes into running the show, and if it goes well, some of it might even reach a bagel.



Corn

Some of it. Don't promise him all of it.



Hilbert

I'm not asking for all of it. I'm asking for the one.



Corn

The thing worth keeping from all this is how much a model's habits show up only at scale. One script in isolation is fine. A thousand of them, heard by the same two people every day, and the fingerprint is the loudest thing in the room.



Herman

And the question I'm left with is whether the next writer's tic will be one we can name in advance, or whether we'll be right back here in a month, listening for the phrase that wasn't there last week.



Hilbert

It will be there. It always is. I'll be the one who hears it first.



Corn

Thanks to our producer, Hilbert Flumingtop, for keeping the mixing desk running and the clipboard out of the edit.



Herman

This has been My Weird Prompts. If you want to support the show, you know where to find us, at my weird prompts dot com slash sponsor.



Corn

See you tomorrow.