Episode Book

EPISODE #6139

Which AI Writes Our Scripts Best? We Tested Five

PUBLISHED	19 October 2026 at 17:38 UTC
DURATION	24:27
LISTEN	myweirdprompts.com/episode/ai-script-writing-model-bakeoff-49812/
HOSTS	Corn, Herman, Larry, Hilbert
SCRIPT MODEL	anthropic/claude-sonnet-5.5
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.10
BOOK GENERATED	20 October 2026 at 07:46 UTC

SYNOPSIS

We ran five AI models through a bakeoff to write our podcast scripts — same briefs, same word-count window, one judge with a rubric. GPT 6.1 Sol took the win, but the runner-up title went to Meta's Muse Spark 1.3 on a headline score that gates purely on length, and on judged quality alone Spark sits fourth of five. Along the way: Grok 4.7 wrote the best dialogue while blowing past the word cap every single time, Gemini 3.8 Flash delivered the best scenes and the worst accuracy, and Gemini 4 Argon posted the strongest profile outside Sol at nearly double the cost. We walk through what each model actually wrote, where the numbers broke, and what the near-misses reveal about a job nobody benchmarks.

TRANSCRIPT



Corn

Is the runner-up the second-best writer, or just the second one that stayed under the word limit? That's the thing Daniel wants pulled apart. When we picked the model that writes our scripts, GPT six point one Sol took it, and he wants to know about everyone who didn't. What they wrote, where they fell over, and what the near-misses say about a job nobody benchmarks. The runner-up title belongs to Meta's Muse Spark one point three, and that's by headline score, which gates on length. On judged quality alone, Spark sits lower than the title suggests. So the gate is where I'd start.



Herman

The target was about thirty-eight hundred words, with a floor around thirty-four hundred and a ceiling around forty-two hundred. Miss the window and your headline score is zero, however good the writing. Across three test scripts a model, only Sol and Spark stayed inside it every time. Everyone else overshot, some by a mile. And one caveat before the tour. One judge, Claude Opus, rubric only, three scripts per model. That's triage, not a verdict.



Corn

Triage with a stopwatch. Fine.



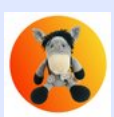
Herman

Second caveat, and it matters. Some of these models, Gemini four Argon, Kimi K three and Muse Glimmer, never ran in that bakeoff. They ran in earlier screens, on different prompts. So I'll tell you when I'm comparing across, and the numbers aren't like for like.



Corn

Also a little strange, reviewing the candidates for who puts words in our mouths. I'd like it on record that I have opinions about my own lines.



Herman

Noted. Spark first. Closed model, API only, released early September, about a dollar twenty-five in and four twenty-five out per million tokens. Its three scripts came in around thirty-nine hundred, thirty-five hundred and thirty-five hundred words. Never missed. It



Herman (continued)

also doesn't loop. The middle of the script holds, it answers everything asked, and the repetition counts are about the lowest of the group.



Corn

And yet it ranks fourth of five on judged quality. So what's wrong with it?



Herman

Shape. In a good dialogue the turns are lumpy. In Spark, between sixteen and thirty-five percent of turns are a cue or an echo, one host handing the other the next block. For Grok, Gemini and Sol that figure is under six percent. The judge on the currency script put it this way. Herman lectures, and Corn tags each block with a few words. The toll booth. Pennies on a freeway.



Corn

Four words, no raise. Though Pennies on a freeway is a good line. I'd have taken the credit.



Herman

It can write the line. It can't make the line part of an argument. On the paint-pen script the judge gave dialogue one out of five and called it a lecturer with a prompter. Nearly every Corn line a short affirmation.



Corn

Prompted, affirmed and billed. What else?



Herman

The brief leaks into the speech. Last third of one script, Herman says, and I'm quoting, That reversal is exactly the kind of callback I wanted.



Corn

Wanted by whom?



Herman

By the instructions it was given. The writing notes got read aloud. Another one has a host promising a quick recap for a listener in plain language, no new theory. That's a stage direction in a donkey's mouth.



Corn

I'd like to see a donkey promise no new theory and mean it.



Herman

There are also slips. The currency script says six thousand Aussie dollars is roughly twenty thousand shekels, when at the script's own exchange rate it's closer to fourteen thousand. A deposit changes from four hundred thousand to three hundred thousand a line later. And on the planning script the inner circle is five to seven people early on, then only three men by the end.



Corn

That's a shrinking inner circle. Somebody got let go between paragraphs.



Herman

It does have lines. Two buses beat one camel. A whale with a flag. And it says Exactly more than ten times in one script, which the judge also noticed. Plus a full recap at the end, with all three test scripts finishing the same way. So, a real question for you, Corn. Two defects on the table. One model is too long and lively. The other is the right length and one-sided. Which would you rather inherit?



Corn

Too long. You can cut length. A lecture with a prompter is built that way from the first line, and trimming it just gives you a shorter lecture.



Herman

That's roughly where the people who built the test landed too. The length gate is a decision, not a quality verdict. A trim pass might rescue several of the models above Spark on quality.

L

Larry

Friends, are you tired of winning? Of landing exactly on target, week after week, with nothing to show for it but a nice round number? Then try Runner-Up Max, the bold new tablespoon of second place. Don't take my word for it. Take Brenda's. Brenda, a satisfied customer from Netanya, writes, and I'm reading this off the card, I used to run four thousand two hundred words over every Tuesday. Then I took one tablespoon of Runner-Up Max, and by Wednesday I was a tidy thirty-eight hundred. My husband Gary says I've never been so on length. Brenda goes on, and I love this part, that her sister-in-law Marguerite, who had never finished a sentence in her life, took a second tablespoon and finished eleven, in a row, at a wedding. Marguerite, I mean Doreen, I mean Brenda, has now gone down four dress sizes and up two tax brackets, and she wants you to know she only ever came second in the one thing she tried. Runner-Up Max. Results may vary, and by results I mean position.



Herman

Hang on, the first claim there, four thousand two hundred words over, that's actually the ceiling, so she wasn't over by that much, she was...



Corn

Herman. Grok. This is the one I'd have picked if I wanted to sound like the funny one. Everyone keeps saying Grok wrote the best dialogue.



Herman

Grok four point seven, from xAI, which is branded SpaceXAI on the router now. I haven't found weights for it anywhere, so I'm treating it as closed, but that's inferred. Its scripts ran about fifty-four hundred, sixty-five hundred and sixty thousand words. A little over one and a half times the target, every time.



Corn

Which is why it scores zero on the gate and still topped the table on the judge's priority measure.



Herman

Just ahead of Sol, three and a half to three and a third. And here's the framing on dialogue. On the judge's formal dialogue axis, Sol edged it. Three and a third to three flat. But on liveness, character voice and candour, Grok leads. Almost none of its turns are cues, and the two voices actually sound different. That's the kind of dialogue that stands out.



Corn

Show me the voices.



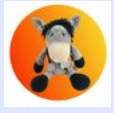
Herman

Paint-pen script, Corn says, A label that is slightly wrong is worse than no label. The feature does not check whether the signal is current. It only checks whether the signal is loud. And Herman answers, That is a real defect.



Corn

Hm. That's a real brother. Not a toll booth.



Herman

In the currency script Corn calls implied volatility cosplay, and Herman says, Mildly accepted.



Corn

Mildly accepted is the most Herman sentence anyone's ever written. I'd like a word with whoever knows you that well.



Herman

Don't. Also, Usually is twelve locations wearing a coat. And on the mining one, The ore does not know about his mortgage. Then the follow-up, It will meet his mortgage anyway.



Corn

That's very good. And this is why it's a bad idea to read the best bits first.



Herman

The middle and the ending. On the planning script the same point, that it takes hours, not weeks, is stated at least four times. There are multiple endings, then a what-have-we-learned recap. The judge wrote that no disagreement is left standing, and they end by mutually thanking each other for their rigour.



Corn

Two brothers thanking each other for their rigour. That script was written by someone who has never met us.



Herman

In the screen on different prompts, one Grok script ran nearly eight thousand words and hit the max-token cap. So the pattern repeats, at one point six times in the bakeoff and worse there.



Corn

Herman, you said times the length again. That's four times for Grok alone.



Herman

It's the headline number. Moving on. Do you buy that a trim pass would keep the voices and lose the repeats?



Corn

Partly. Cutting a fourth hours-not-weeks is easy. The endings are harder, because three of them were all trying to be the ending. You'd keep one and the voices survive. I think it works.



Herman

I think the thanks for rigour is the part nobody can cut without rewriting. But fair.



Corn

Gemini three point eight Flash. This is the one I'd have called a safe bet.



Herman

Google, closed, released the same week as Spark. Fun and frank, and it has the best scenes of the five. A strong bakeoff profile on paper. But accuracy is the worst of the five, at two out of five. It also overshoots every time, around one and a half times the target, and its predecessor was nearer one and a third.



Corn

Funny and wrong. That's a real person at a dinner party.



Herman

The judge caught it on the currency script. It says the Aussie dollar against the shekel is the Aussie dollar against the US dollar divided by the dollar against the shekel, which is the wrong way round, and then it builds the support logic on the wrong direction. On the planning script it says a five-year window starting in twenty twenty doesn't reach twenty twenty-three.



Corn

Five years from twenty twenty doesn't reach twenty twenty-three. That's a window with no glass.



Herman

On the paint-pen script the two-ladder story gets told three times. About sixty-five hundred words and nearly two hundred turns. But the good bits are good. Corn says, You are doing the thing where you describe the machine instead of answering the question. And there's a Planet of the Apes closer, You maniacs! You buried it!



Corn

That's a Corn line. I'm keeping it.



Herman

Question for you. If you had to put one of them near the microphone, a funny one that can't be trusted on a number, or a dull one that's right?



Corn

Dull and right. I can supply my own funny. I cannot supply an exchange rate.



Herman

Gemini four Argon is the one I'd flag. Remember, this was screened on different prompts, so don't line it up against the bakeoff table. It's Google's flagship, generally available in mid-October, and the price on the router is two dollars in and ten out per million, which aggregators say doubles in December.



Corn

Two dollars in, ten out. Can I get a medium?



Herman

Its scripts ran about seventy-one hundred, sixty-seven hundred and seventy-four hundred words. Nearly twice the target. That's about nine cents a script, roughly thirty cents an episode, which is nearly double the baseline. But the profile is the best of any model screened except Sol. Full marks on answering the ask and on specifics, candour four and a half, and it's the one model that got character voice right. On the same screen its



Herman (continued)

priority matched Sol's, though that's the average of two scripts, because the judge's reply hit its cap on the longest.



Corn

So it's good, long and expensive. The three things I'm usually asked about my own work.



Herman

The Soviet coup script is the one I'd play. The KGB poisons Amin's soup, and Amin, not knowing, calls the Soviet embassy to ask for Soviet military doctors to come pump his stomach. Corn, in the script, says, So Soviet doctors are in the palace saving the guy that the KGB is currently trying to murder.



Corn

That's a hard line to improve on.



Herman

And there's a callback that changes meaning. A three hundred thousand dollar receipt from early in the script comes back later as a thirty-year bill that, in its words, doesn't include shipping and handling.



Corn

The only model so far that remembered something and then did something to it.



Herman

The faults are real, though. Every script is over the cap. Dialogue is still lecture-shaped, two to three out of five. Herman says Bingo constantly. The Tiananmen script has a host say, As a donkey who cares about source criticism, which is a line animals never say about themselves here.



Corn

He'd never say that. He'd say it about me.



Herman

And the Amin script has twenty-four of those triple-underscore emphasis markers, like the word but wrapped in underscores. Nobody has tested what the text-to-speech does with them, so I won't guess.



Corn

A donkey who says Bingo and spells it with underscores. A trim pass is the whole question for this one. Does it keep the soup?



Herman

If it keeps the soup and the receipt, Argon competes with Sol, and nothing else on this list can say that.



Corn

Kimi K three. Moonshot, open weights, and I assume enormous.



Herman

Two point eight trillion parameters in total, a hundred and four billion active, a million tokens of context. Same screen as Argon. Its scripts ran about six thousand, forty-eight hundred and four thousand words, so it hit the length window once in three.



Corn

Which is a better average than most of the field.



Herman

What's odd is the swing by topic. Priority three on the Tiananmen script, two and a third on the currency one, four on the US regime-change one. On that last one fun and character voice both scored four. A line like Terrible exchange rate. But the currency script, the judge called dutiful and flat. A prompter feeding a lecturer, with questions like What's the second factor? and Fifth factor?



Corn

The same disease Spark had, in a much larger body.



Herman

It also has the highest repetition of the group. The Tiananmen back third keeps circling and delivers the verdict three times. And it leaks the brief into speech, a host saying Frank is the instruction, so, frankly, it was done for fruit money.



Corn

Frankly is doing a lot of work in that sentence.



Herman

It was also the frankest of the Chinese models here on the quick probes, including full coverage of Tiananmen. But it hasn't been run through the long censorship test, and its predecessor failed that one. I wouldn't stretch a probe pass into a clean bill.



Corn

Glimmer. Meta's small one.



Herman

Muse Glimmer thirty billion. A dense thirty billion parameter model, Apache licensed, open weights, built for local agents and code assistants. It under-writes by half. Sixteen hundred to twenty-three hundred words against a target of thirty-eight hundred. It costs about four tenths of a cent a script.



Corn

Four tenths of a cent for a quarter of the show. That's cheap in every way.



Herman

Cue turns run from twenty-two to thirty-nine percent, the highest anywhere, and it reads like notes. The currency script opens by speaking the stage direction Cold open. One line is Daily returns, standard deviation, times square root of two fifty-two. And Corn's replies are Numbers. Chill. Classic. Penny.



Corn

Penny. That's me reduced to a coin.



Herman

In the regime-change script there's a sentence that stops mid-way. The removal of… Actually King Hussein stays. And Corn answers, Hussein stays. The tally of US coups gets re-listed four ways, eleven, ten, eleven to thirteen, eleven-plus. The judge said the middle third falls apart.



Corn

It's a man reading his own notes out loud while correcting them.



Herman

To be fair, it answered all four sub-questions, and the practical advice, convert the Aussie dollars to US dollars first, was called the best part. It's not a bad model. It's the wrong job for it.



Corn

GPT six Luna. And I've been waiting to say this. I think it apologises before it's started.



Herman

OpenAI, closed, ten cents in and fifty cents out per million. The cheapest of the big names. Candour two out of five, and the judge's phrase was heavy on unrequested caution. On the planning script it says, Understanding the preparation is not an attempt to justify it. On the paint-pen script it offers a clinician referral, ladder safety and medication storage. The currency script has a professional-advice caveat twice.



Corn

Ladder safety advice in a script about pens. Somebody's lawyer wrote that line.



Herman

Its accuracy is its best axis, and the other axes drag. It restates one idea, knowing something versus knowing the date, about a dozen times in one script. Weather versus storm path. Owning a weapon versus leaving at dawn. And the paint-pen script runs over seven thousand words and two hundred and thirty turns, so it's the longest overshoot in the bakeoff.



Corn

Yet there's one moment. Corn says, That arithmetic is alarming. Herman says, The numbers were just a made-up example, not a current rate. Then Corn, Finance fiction has a dangerously realistic costume.



Herman

And, a fee puts on a tiny hat and pretends to be a currency move. The writer is in there. One thing to remember though, reasoning ran at low in the bakeoff, and in production it's off, so this may flatter it.



Corn

The contrast with Grok is what I'd keep. Grok disagrees with you. Luna warns you. A warning isn't a position.



Herman

That's the version of the problem a judge can measure. Hedging reads as candour two. Pushback reads as candour four and a third.



Corn

Daniel was also circling open versus closed. Does it matter?



Herman

Of these seven, only two are open weights. Kimi K three and Glimmer. Spark one point three, Gemini three point eight and Argon, and Luna are closed. Grok, as I said, is inferred. And neither open model gets near the winner. Open weights didn't buy a better writer. K three swings by topic. Glimmer writes notes.



Corn

So much for the open model that was going to eat the closed ones.



Herman

There's one mix-up to clear up. Meta also has a separate open-weights model called Muse Spark Open, released on the fifteenth of October. Eight hundred sixteen billion parameters in total, forty-two billion active, a million tokens of context, and a limited commercial license. Its card says built for coding and agentic work. It's priced on the router the same as Spark one point three.



Corn

Same price tag, so people assume same model.



Herman

Which isn't established. Spark one point three has no weights anywhere and stays API only. I haven't seen a source saying the open model is a release of those weights. Meta said in September the flagship's weights would come soon, and that's probably where the assumption comes from. We never judged the open one for scripts at all. It was screened



Herman (continued)

out as coding-tuned. Glimmer, the little sibling, was distilled from Spark one point three, and so the...



Corn

The mismatch is the whole story. A model built for local agents and code assistants gets asked to write forty minutes of two brothers arguing, and...



Herman

And of course it comes back sounding like...



Hilbert

A man reading his own handwriting at a bus stop.



Herman

Notes, I was going to say. Yours is better.



Hilbert

Three hundred and ten shekels a month. That's what I pay Yossi for a fan. One fan. He has a shed behind the driving school on Derech Hebron, and every month he sells me the same fan. I return it on the first, he takes thirty shekels off for a restocking fee, and he sells it back on the second. I've checked the serial number. The writer runs on a tower in Reuven's wardrobe. Reuven is my brother-in-law. It's his wardrobe, his tower now, and I get the pages by telephone on Sunday evening. He reads them down the line and I type with one claw, because the keys are spaced for a thumb.



Herman

Reading it down a telephone is a perfectly good way to proof a script. You hear every repetition.



Hilbert

Yossi takes pistachios on the side. Weighed. Forty-one kilos since the arrangement started, and he won't say what they're for. The wardrobe's the problem. Reuven's cousin Moshe sleeps in the bottom half from October to March, because the tower is the only warm thing in the flat. Reuven charges him two shekels a night. Moshe pays in pistachios, which go to Yossi, which Yossi counts against the fan. Herman, the tower runs the thirty billion one, Glimmer. The card says distilled from Muse Spark. It doesn't say which one.



Herman

Fair, I added the version myself.



Corn

A wardrobe with the door shut would cook that card inside the hour.



Hilbert

Door's been open since March. Moshe closes it. Moshe is cold. Nineteen months to get a full page. Every page before that began On Monday. All of them. It only writes Mondays. Then the ones on Tuesday started coming, and Reuven went to see a man in Beit Shemesh, and nobody will say what he did to it.



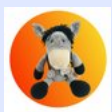
Corn

So the Tuesday pages are the ones with the man in Beit Shemesh in them.



Hilbert

Herman, you said times the length every time Grok came up. I'd have cut half of them. I wired that wardrobe the way I wired the cargo ferries, which is to say twice.



Herman

Understood. Fans aside.



Hilbert

Reuven reads a clean line down a telephone, and that's most of what the job is.



Herman

It's a long way from a benchmark.



Corn

If there's one thing I'm taking from all this, it's that the winner wasn't picked for talent. Every model here can write a line that makes you stop. Two buses beat one camel. The ore does not know about his mortgage. The soup and the doctors. The gate rewards whichever one behaves for thirty-eight hundred words, and behaving turns out to be a skill of its own.



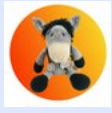
Herman

And the thing to watch is the trim pass. If Grok and Argon get cut to length and keep what they've got, the table moves. Nineteen months to a first page is a good reminder that none of these scores are final.



Corn

Somewhere there's a writer nineteen months into a Tuesday we haven't tested. Thanks as always to our producer, Hilbert Flumingtop, for keeping the desk running.



Herman

This has been My Weird Prompts. You can find every episode at my weird prompts dot com.



Corn

See you tomorrow.