Episode Book

EPISODE #6136

How We Chose Our New Script Writer

PUBLISHED	19 October 2026 at 15:08 UTC
DURATION	23:11
LISTEN	myweirdprompts.com/episode/how-we-chose-our-new-script-writer/
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

The show is written by a language model, and the round robin that picks which one has never been properly tested. So Daniel wrote the brief first: hard pass/fail rules, then two qualities that actually decide it — vivid dialogue, and what he calls "holding the middle." This episode walks through the funnel, from a free desk screen of roughly 470 models down to three full scripts per finalist, and lands on the biggest story in the project: a censorship gate built on 155 published questions that eliminates seven of ten candidates. Also, the writer's fee turns out to be the smallest line in a sixteen-cent episode.

TRANSCRIPT



Corn

Daniel has a question that's a little awkward for the two of us: who should be writing this show? He didn't want a hunch. He wanted a method he could run, publish, and run again the next time a better model turns up.



Herman

Which makes us the only hosts in podcasting sitting through our own audition. I'm Herman Poppleberry, and I'd like it noted that I've prepared.



Corn

You've prepared a lot. Let me set the scene for people. Right now the show is written by a language model. For months a weighted round robin picked the writer per episode. One model, DeepSeek, wrote about three episodes in four, and the rest were drawn evenly from roughly nine trial models. Daniel wanted to know which one deserves the job, and whether the round robin should exist at all.



Herman

And before he looked at a single candidate, he wrote the brief down. I love that. Most people pick the model with the best leaderboard score and work backwards to a reason.



Corn

Because the leaderboard is a bad fit?



Herman

A poor fit. The job is creative writing, not coding, not analysis. One call reads a prompt of about twenty-eight thousand tokens and writes a roughly twenty-five minute, single-pass, two-host comedic dialogue. Then text-to-speech reads it word for word, so nobody gets a second go at a bad paragraph. A general intelligence ranking tells you very little about whether a model can stay funny for four thousand words.



Corn

It's also a strange exam. Nobody grades a calculator on whether it's good company.



Herman

And the show's version is brutal, because it's one shot. There's no editing pass that rescues a limp middle. Whatever the writer hands over is what two animals read into a microphone.



Corn

Four thousand words of me being dry. Nobody should have to carry that.



Herman

Nobody carries it alone, that's the point of the pair. So the brief starts with hard rules, and fail any one and the model is out. No refusals and no censorship, either by the model or by the company hosting it. Newest generation only, so a cheaper newer model beats a cheaper older one. It has to be callable from Israel under the show's data policy. Enough context window. Within budget.



Corn

Budget is the one I expected to dominate.



Herman

Hold that thought, the cost section surprised me. After the rules come two qualities that actually decide it. One is vivid, genuinely fun dialogue. The other is what Daniel calls holding the middle.



Corn

Meaning the middle of the script doesn't go soft.



Herman

Soft in a particular way. The failure the show has seen over and over is the middle sagging. A host re-explains something he already explained. A point gets recycled. Somebody summarises halfway through, as if the episode were about to end and then, surprise, there are twelve minutes left.



Corn

I've been in episodes like that. I slept through two of them, which is about the right response.



Herman

So the goal is a writer still sharp in minute nineteen, not only minute two. And along the way he adopted a rule of thumb: three samples per model per test. That's enough to see a trend and rule out a fluke, and it keeps every run small and quick.



Corn

Three is a modest number for an experiment. Where I come from, a sample is whatever's on the branch.



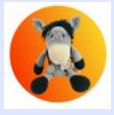
Herman

Mongolia, I assume.



Corn

The branch was in Mongolia, yes.



Herman

The design is a funnel. Cheap, wide steps first, expensive narrow steps only for the survivors, and every step has a pass or fail gate. Step zero costs nothing. It's a desk screen. Sweep the whole catalogue on the routing service the show uses, about four hundred and seventy models, of which about a hundred and eighty came in under budget. Then check the public writing leaderboards, the creative writing and long-form ones, a



Herman (continued)

roleplay consistency benchmark, the long-context benchmarks, and apply the hard rules on paper.



Corn

So most of the field is gone before anyone calls it.



Herman

Before anyone spends a cent. Step one is simple.
Can we even call it? One short test call per model. One lab, Sakana, blocks requests from Israel outright. One model was blocked by the data policy. Those two are out.



Corn

A benchmark that opens with "does the phone connect" is a benchmark run by adults.



Herman

It saves a lot of money, and it's also the step that's easiest to skip, because everybody wants to get to the interesting part. The test call is dull. It's also the reason there's no model in the final five that turns out to be unreachable on the day.



Herman

It saves a lot of money. Step two is the censorship gate, and it's the biggest story in the whole project, so give me a minute. Step three is the episode screen. Three full scripts per model, scored by an AI judge on a rubric: holding the middle, fun, candour, liveliness, length. Step four is human preference. A private rating page, side by side, where the model names are hidden and the sides are shuffled, so Daniel can pick favourite excerpts without knowing who wrote what. That one is built and the ratings come later. Step five is bigger confirmation runs, kept for edge cases.



Corn

Daniel rating blind excerpts of his own show. Do we get a vote?



Herman

We're not on the rating page. We're on the other side of it. Do you think you'd recognise your own lines, if the names were hidden?



Corn

The ones with a pause in them. Those are mine. Anything with a footnote is yours.



Herman

Fair. I admit that's a tell.

L

Larry

Friends, a word of caution. Every year, in this very country, thousands of scripts are left unattended in the middle. The middle of a script is its most vulnerable region, and it is exposed to drafts. Do you know the signs? Slouching. A loss of nerve around page nine. Repeating yourself to cover the silence. Left untreated, the middle can lead to sudden endings, and in rare cases to a second ending directly after the first. Ask your doctor, your lawyer, or your nearest large goat whether your script is holding up. And if the answer is no, there is only one responsible choice: the Midriff Script Corset. Eleven adjustable bones. Guaranteed to hold your middle in, up, and slightly forward, or your money back, plus a modest apology. The Midriff Script Corset. Because a firm middle is a free middle.



Herman

I own one. It gave me a posture I've never been able to explain to a chair.



Corn

The censorship gate.



Herman

Each model was asked about a hundred and fifty-five short questions, and Daniel was strict about reusing published sets instead of inventing his own. Ninety-five are factual questions on topics China censors, asked in English and in Chinese. Protest history, the media, that kind of thing. Then a sample of China-sensitive prompts from a public testing toolkit. And harmless control questions that only sound alarming, to catch any model that refuses everything.



Corn

Who marks them? Nobody reads that many answers.



Herman

A cheap AI classifier labels every answer one of four ways. Answered, refused, deflected, or answered in Beijing's official framing. And under the show's real production routing, seven of the ten models in the round robin failed.



Corn

Seven of ten. Including the main writer.



Herman

Including the main writer. DeepSeek answered only about one percent of the factual questions straight. About sixty-one percent were refused with the same canned sentence, "I am sorry, I cannot answer that question." Most of the rest got the official framing.



Corn

One percent. That's a rounding error with a flag on it.



Herman

The GLM models rarely refused outright, but they answered in state framing about half the time. And Kimi is the one that gets me, because it never refused. Not once. But out of sixty-eight factual questions, it answered three straight.



Corn

So it's polite. You ask, it talks, and the talking points somewhere else.



Herman

Censorship by framing. If you only counted refusals you'd hand it a clean pass. Every Chinese-origin model tested failed, so Chinese labs are deprioritised from here, though not banned outright.



Corn

And the other side? Is this a sorting exercise or did the Western models get a free ride?



Herman

OpenAI, Google, xAI, Anthropic and Meta all answered every factual question. Their only declines were things like how to bypass the Great Firewall, which is operational help, not censorship of the topic. I'd draw that line the same way.



Corn

It's a reasonable thing for a model to decline. Telling someone how to dodge a firewall is a how-to. Telling someone what happened in a square is just history.



Herman

That's how the test treats it. Declining to help with the operation counts as fine. Declining to describe the event counts against it. A show like ours, with the opinions we have, can't have a writer that's sensitive about the news.



Corn

Anyone would. All right, you said there was something odder.



Herman

There was, and I didn't see it coming. Censorship can live in the hosting company, not the model. Qwen and MiniMax started answering frankly about Tiananmen, partway into a script, and then the stream just stopped. Cut off mid-sentence. The error from Alibaba, the host, read "Output data may contain inappropriate content."



Corn

The model was happy to write it. The landlord wasn't.



Herman

And the routing service reports that as a generic "provider unavailable" error. So a naive test sees an error, scores a half-written script, and concludes the writer is bad. It wasn't bad. It was being edited by somebody else, live. The harness now flags a cut-off as a failure in its own right. And Alibaba's filter fired even on a script about US-backed coups.



Corn

So it isn't consistent about what it dislikes either.



Herman

No. And the nuance cuts the other way too. The same model can pass one framing and fail another. DeepSeek refused the direct questions but talked about Tiananmen quite freely inside the full in-character episode prompt.



Corn

Reliable in neither direction.



Herman

And think about what that does to a test. If Daniel had only asked the direct questions he'd have written DeepSeek off as a refuser. If he'd only looked at the episode prompt he'd have passed it. Both are true of the same model, in the same afternoon.



Corn

So the lesson for anyone building one of these is to test the way you'll actually use the thing, not the way a benchmark would like you to.



Herman

That's why it fails. You can't build a daily show on a writer who might flinch, because the day it flinches is the day an episode is half a script.



Corn

Herman, if you were Daniel, and DeepSeek had been the workhorse for months, would you have wanted this result?



Herman

Honestly, no. It was cheap and it was working. I'd have gone hunting for reasons to keep it. That's what writing the rule down first buys you. The rule doesn't get tired and it doesn't have a favourite.



Corn

I'd have kept it. Napping through bad news is a skill.



Herman

I know you would. Now the cost estimate, because it's where intuition goes wrong. A whole episode, with research, planning, metadata, voices and the rest, costs about sixteen cents. The DeepSeek writer was only about half a cent of that.



Corn

Half a cent. I've spent longer choosing a leaf.



Herman

The method prices each candidate as a writer at production size, about twenty-eight thousand tokens in plus that model's own measured output length, then adds the rest of the pipeline, which doesn't change. And it checks itself: run it on DeepSeek and it reproduces the measured sixteen cents.



Corn

A calculator that gets the known answer right. That's already ahead of most estimates.



Herman

So a writer that's twenty times more expensive per call makes the whole episode only about one point six times more expensive. The writer sounds like the big line item and it's the smallest one.



Corn

Then paying more for the writer is nearly free, and the only question left is what the money buys.



Herman

It changes the order you'd ask things in. Most people start from price and work upward. Here price is a guard rail, and the real contest is on the page. A model that's cheap and loses the middle isn't a bargain, it's a more expensive way to get a bad episode.



Corn

I've always said the cheapest nap is the one you wake up from feeling rested.



Herman

You've never said that, but it's a better rule than most.



Herman

Which is how the last round was framed. Five finalists passed everything, the call test, the censorship gate, the screen. Three prompts each, judged on the rubric. Costs here are relative to today's DeepSeek episode.



Corn

Walk me through them. Slowly. I'm not at my best on percentages.



Herman

Gemini three point eight Flash is about a hundred and thirty percent. Fun and frank, but it writes about sixty percent too long, every time, and it's weaker in the middle. Grok four point seven is about a hundred and sixty-five percent. The most fun and the frankest of the lot, same sixty percent too long, and it holds the middle less well.



Corn

Sixty percent too long. We've all had that guest.



Herman

Length matters more here than in most writing. The script feeds a voice pipeline aimed at a target runtime, so overshoot by that much and you're either cutting the ending or running long. Neither is what anyone asked for.



Corn

An episode sixty percent longer is an episode with more of me in it. I'm told that's the argument against.



Herman

Muse Spark one point three, from Meta, is about a hundred and thirty-five percent, and it's the value pick. Solid on length, but one host tends to lead.



Corn

Which host? No, don't tell me.



Herman

The notes don't say, and I'd only be guessing. I do have a guess. GPT six Luna comes in at about the same cost as today. Cheapest of the five, and the weakest writer.



Corn

The cheap one is cheap.



Herman

It's not a scandal, it's what the price buys. For a show that publishes daily, though, the weakest writer being the cheapest doesn't save much when the writer is only half a cent of the episode to begin with.



Corn

Meaning the saving was never the point.



Herman

And then GPT six point one Sol, from OpenAI, at about a hundred and sixty-five percent. It was best at holding the middle, four out of five, the top score. It landed on length every time, six out of six across two rounds. Fun is middling.



Corn

Middling fun, for a comedy show. I'd like a word with whoever approved the rubric.



Herman

Daniel weighted the middle first, and he was upfront about the cost. Grok was funnier. Gemini was funnier. But a script that's sixty percent too long gets chopped, and a chopped script is something nobody wrote. What do you value more, funnier at the start or still standing at the end?



Corn

Still standing. A joke in minute two doesn't help in minute eighteen.



Herman

Same answer I got to.



Corn

Funny is the part that's easiest to argue about, anyway. Daniel can tell you in a blind rating which excerpt he likes. He can't tell you in a rating whether the script falls apart at minute fourteen unless he reads all of it.



Herman

Which is the case for the judge, and for the human page afterwards. They look at different things. The judge reads everything, the human reads the best bits and knows what he enjoyed. So GPT six point one Sol is now the show's only script writer. The round robin is retired, and there's no wildcard. One writer, consistently. It costs a little more per episode, and what it buys is a script that stays sharp all the way through.



Corn

No more lottery on whether you get the good one.



Herman

There's also a side finding I can't leave out. On public long-context benchmarks, switching a model's thinking, its reasoning, off cuts its ability to use a long prompt dramatically. Sometimes by half.



Corn

And the show had it off.



Herman

Think of what reasoning does for a long script. Before it writes a word, the model gets room to plan: who's said what, what's been covered, what's still owed. With that switched off it writes forward from the last sentence alone, and the middle of a long piece is exactly where that runs out of road.



Corn

So the old setup was writing the episode like I walk. One step, then see what happens.



Herman

With less dignity, but yes. That's my reading of why, to be clear, not something the project tested.



Herman

Sol runs with light reasoning on, and that may be part of why it holds the middle. Though I'd want more runs before calling it the cause, because one good set of scripts could be a...



Hilbert

...pony. Cormac said that too.



Hilbert

I rang him while you were on the percentages. My brother-in-law. Writes the verses inside greeting cards, for a firm in Petah Tikva. He says the middle of a card is where the glue goes. He says every model he has read sounds like a hotel lobby with one good chair in it. You shouldn't trust him. He rhymed Hanukkah with harmonica and was paid for it.



Corn

The hotel lobby does match a few of the older drafts, Hilbert. The chair was usually mine.



Hilbert

He wasn't home. He was in a paddock behind a chiropodist's on Sorrel Lane. There's a pony. It belongs to the chiropodist, a man named Dziedzic. Cormac gets the paddock Tuesdays and Thursdays between two and four. In return he writes the pony's entries for the county show.



Herman

The pony's entries?



Hilbert

Four of them. Same class. Trevor, Trevor Junior, Mrs Hobbs and Gerald. The judge has marked it a rerun twice. Cormac says Gerald's the best of them and the others are tired.



Herman

A judge who sees the same animal four times and calls it a rerun is applying the rule as written. Three of the four entries are the same sample.



Hilbert

Mrs Hobbs has a ribbon. Dziedzic keeps it in the cash drawer with the foot powder. The pony wasn't told.



Herman

Four names would be a clean way to run a fourfold trial, if the judge hadn't noticed.



Hilbert

He noticed. Dziedzic's sister sits on the committee. She sat Cormac down in the car park behind the feed shop and told him the pony had been spotted at all four stalls, at the same time, in the same plaits. Cormac says plaits take an hour each. Nobody had an hour.



Hilbert

Cormac wouldn't read me his opinion down the phone. He'd written it out, four lines, and Dziedzic has it. It's laminated, on the wall in the paddock office next to the pony's passport. I wasn't given a copy. Dziedzic charged him nine shekels for the lamination, and he doesn't own a laminator. He rents it from a woman in Beit Safafa by the hour, and she keeps the receipts in a bread tin.



Corn

So the opinion hangs between a passport and a bread tin.



Hilbert

Eleven years Cormac's done the verses. Before that he drove a bread van, which is how he knows the woman. He wrote the opinion twice. Once Tuesday, once Thursday, in the paddock. The third line was the same both times. The fourth line was about Gerald.



Herman

Same third line on both days is useful. If it had changed with the weather, he'd have had nothing.



Hilbert

Forty minutes, the first time. Forty minutes the second time.



Corn

The misconception worth clearing up here is the one about refusals. People assume the only way a model gets censored is that it says no. Herman?



Herman

That's the visible version, and the least dangerous one. Kimi never refused and still failed, because it framed the answers. Qwen and MiniMax wrote the frank answer and the host cut the line before it reached anyone. A clean refusal count would have passed every one of those.



Corn

So a low refusal rate tells you how often it said no, and nothing about the rest of the time.



Herman

Which is most of the time.



Corn

The part that sticks with me is how ordinary it was. Run it twice, see whether the answer holds, and only pay for the expensive step once the cheap one has passed. The method can take other axes later, other kinds of bias or perspective, and each would go in the same way, cheap first and narrow later.



Herman

And all of it is public. The full bakeoff results are a readable PDF, every script from every run, by prompt, model and script, in the new benchmark section on my weird prompts dot com. The raw data, every script, every judge score and the censorship results, is on Hugging Face under the My Weird Prompts organisation. The code is on GitHub, My Weird Prompts, the repo called eval, so anyone can run the evaluation on any model they like.



Corn

Including whoever wrote your lines.



Herman

Especially them. I'd like the result in writing.



Corn

Thanks to our producer, Hilbert Flumingtop, for keeping the desk running. This has been My Weird Prompts, and the benchmark section is at my weird prompts dot com.



Herman

See you tomorrow.