Episode Book

EPISODE #6713

Airing Challengers to Test AI Podcast Writers

PUBLISHED	20 November 2026 at 14:58 UTC
DURATION	24:38
LISTEN	myweirdprompts.com/episode/airing-challengers-ai-writers-06842/
HOSTS	Corn, Herman, Dorothy, Hilbert
SCRIPT MODEL	anthropic/claude-sonnet-5.5
VOICES	omnivoice
PIPELINE	V5.36
BOOK GENERATED	20 November 2026 at 21:20 UTC

SYNOPSIS

A five-part technical series tested whether rival AI models could replace a baseline scriptwriter when evaluating search primitives. Across fifteen drafts from Qwen 4 Max, Gemini 4 Carbon, Gemini 4 Hydrogen, and Kimi K3.1 Pro, zero scripts failed Claude Opus 5.5's evaluation, though every model ignored topic boundaries and trampled episode fences. By design, challengers aired regardless of baseline scores, exposing how research packets shape jokes and why negative prompt constraints regularly fail against heavy source material.

TRANSCRIPT



Corn

A script costs about two cents to write. The voices that read it cost many times that. So if you want to try a new writer, the cheap part of the experiment is the writing, and almost nobody uses that to their advantage.



Herman

We did, today. Daniel wanted the results of the trial Claude supervised on the five-part series about building search from scratch, the one that just finished airing. Fifteen drafts, five models, eighty cents of text in total.



Corn

And you can already tell me the winner.



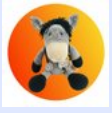
Herman

There isn't one, and I'll defend that. Nobody has called a winner, including the judge.



Corn

Then I'd like to hear how you run a trial with no winner. Start with the setup.



Herman

The series is five parts on the primitives of search: the inverted index, BM twenty-five, the gap between keyword and semantic, embeddings by hand, then hybrid search and how you evaluate it. The test corpus is this show's own archive, six thousand and some episodes of transcripts. For each part, the same prompt and the same series hand-offs went to several writers as script-only jobs. No voices get rendered until a draft is picked.



Corn

Which writers?



Herman

Qwen four Max drafted all five parts. Gemini four Carbon took two, Gemini four Hydrogen took one, and Kimi K three point one Pro took two. And GLM five point three, the live writer, drafted every part as the baseline, because without it there's nothing to beat.



Corn

So I'm being read out by one of the contestants.



Herman

Whoever is writing the show on any given day is on that list, yes. Quite likely you.



Corn

Comforting. Who judged?



Herman

Claude, Opus five point five, one judging pass per part. It read each raw draft whole, before anyone edited a word, and wrote a verdict plus a few lines naming specific lines. Pass means it airs with light edits. Marginal means it airs only with heavy edits. Fail means it doesn't.



Corn

Fifteen drafts. How many failed?



Herman

None. Six passes, nine marginal, zero fails.



Corn

That's either very good news or a very generous judge.



Herman

The breakdown by part is more useful than the total. Part One, Qwen four and GLM passed, Kimi was marginal. Part Two, Qwen four and Carbon passed, GLM was marginal. Part Three, everybody was marginal.



Corn

Everybody. On the one about the gap between keyword and semantic.



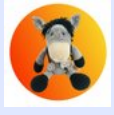
Herman

It's the hardest of the five to do without wandering. Part Four, GLM passed and Qwen four and Hydrogen were marginal. Part Five, Qwen four passed and GLM and Carbon were marginal. So GLM was pass, marginal, marginal, pass, marginal, and Qwen four was pass, pass, marginal, marginal, pass.



Corn

A pass-and-marginal checkerboard. That's the most honest chart I've seen from an AI evaluation.



Herman

Yes, the board shows no pattern you could build a decision on. Now the rule that changes what the experiment is. Daniel's instruction was to air challengers, not to pick the best draft. The first challenger draft that's good enough, a pass, or a marginal whose faults an edit can fix, goes to the voices straight away. It isn't held back to compare against later drafts.



Corn

So airing is the evaluation.



Herman

Airing is the evaluation. And the corollary is that GLM never aired in this series. Not because it lost. By design it only airs as a fallback if no challenger can be rescued.



Corn

So the baseline sat the whole thing out.



Herman

It drafted every part and aired none of them, yes. And if two challengers were both good enough, the one that hadn't been heard yet got the slot. That's why Part Two went to Carbon even though Qwen four also passed it, and Part Five went to Carbon even though Qwen four's Part Five was the pass and Carbon's was marginal.



Corn

Airing a marginal draft over a pass. That sounds like a thumb on the scale.



Herman

It was, openly. Daniel said he was eager to give Gemini four a go, so the trial waited for Carbon's drafts on those two parts and aired them as long as an edit could save them. The point isn't to find the best script for each prompt. It's to get new writers on air so he can hear them.



Corn

Fine. Then Claude edits the draft that gets picked.



Herman

Turn by turn, and the edit is submitted under the name of the writing model, so the credits name the challenger, not Claude. The edit goes in the log as its own row with no verdict. Claude doesn't grade its own work. Daniel grades it along with everything else.



Corn

Which is where I'd start poking. The judge knew who wrote which draft.



Herman

Yes. Not blind.



Corn

One judge, and it's an AI from one lab judging models from other labs. Then the same family edits whatever airs.



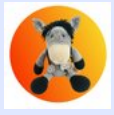
Herman

All fair. The scorecard design episode we did, Building a Scorecard for Our Own AI Writers, lists judge self-preference as a threat. So does the written-up benchmark paper. I'm not going to wave it away, so I'll list the rest. One prompt per part. Each draft ran its own research, so the writers weren't handed identical facts. And what Daniel hears on air is draft plus edit.



Corn

So when Daniel says he liked Part Three, he's grading Kimi and Claude's red pen together.



Herman

Yes. Kimi's Part Three went from about six thousand one hundred words to forty-six hundred in the edit. That's a quarter of it gone.



Corn

Our target was what, thirty-four hundred to four thousand?



Herman

Roughly. On length, GLM averaged about forty-three hundred words a draft, the nearest to the brief. Qwen four averaged about five thousand, running one and a quarter to one and a half times long on Parts Two through Four. Kimi averaged about sixty-three hundred.



Corn

Kimi wrote a double episode and called it a draft.



Herman

Both Kimi drafts came in at about one point six or one point seven times the target. The two Gemini four drafts sat near forty-four hundred, and the judge had no length complaint on either.



Corn

Does the judge say anything good, or is it a list of crimes?



Herman

Plenty good. Let me read some, because the best part of this exercise is the actual text. Qwen four Max, Part Two, the BM twenty-five episode, Herman's line: "A term appearing in half the corpus is worth one bit. Literally one bit, log base two of two. The unit isn't a metaphor, it falls out of the formula."



Corn

That's a good one. A formula making a joke.



Herman

And the same draft on the benchmark everybody uses to test retrieval, where the relevance labels were pooled from the systems being compared. Qwen four, Part Two: "Graded on homework where the answer key was written by the student who's being compared against."



Corn

Harsh, accurate and fits on a sticky note.



Herman

Would you have cut it?



Corn

Never. Cut the Herman in front of it, maybe.



Herman

Qwen four also gave you lines in Part Three. You, on the idea of a neural network used to build a keyword index: "A neural network whose output is a nineteen-eighties data structure." And on averaging meanings: "It's the average of a river and a mortgage."



Corn

Mine, but slower.



Herman

I'm not sure that's a defence.



Corn

What about the others? You've been very Qwen-heavy.



Herman

Fair. Kimi, Part Three, where the judge said the real co-host play showed up. Corn: "Same religion, different plumbing." Herman: "I am keeping that." Corn: "You may not."



Corn

That's more me than I expected. Refusing you something small, for no reason.



Herman

The point underneath is real, too. Old latent semantic indexing and modern embedding models share a hypothesis, that words keep company, and only the plumbing differs. The judge called Kimi the most conceptually sharp of the lot.



Corn

What did Gemini four sound like? It's the newest name here.



Herman

Carbon wrote the cleanest challenger draft, Part Two. The mechanics were sound, with worked numbers on how BM twenty-five's saturation constant behaves: the first mention of a term gets you about forty-five percent of the ceiling, and five mentions get you past eighty percent. And when the draft reaches how the constants were chosen, Corn says: "Wait. They eyeballed it?"



Corn

They did, didn't they. Half the field's famous numbers are somebody's guess that stuck.



Herman

Your line there is the reaction the point needs, which is the judge's complaint against so many drafts, where Corn just nods.



Corn

And the baseline? It never aired. Let me hear it at least.



Herman

GLM, Part One, you, on how long a naive search takes across the full archive: "Eight and a half minutes. That's not a search, that's a deposition." And Part Four, on silent failures in an embedding pipeline: "Nothing in this pipeline screams. Everything whispers."



Corn

Those are good. I'm annoyed at how good.



Herman

There's one more I find the most interesting, and it's about how much of a joke belongs to the model versus the research. Part Five, the hybrid episode. The research turned up that the standard constant in reciprocal rank fusion was set at sixty in a pilot investigation on one collection and never re-tuned, and it's now the default in a stack of products. Qwen four's Corn says: "Seventeen years of infrastructure resting on a shrug." And Gemini four Carbon's Corn, independently: "Seventeen years of infrastructure on a placeholder."



Corn

Same joke twice. Two models, two research runs.



Herman

Same fact, same joke shape. The punchline was sitting in the material. A decent chunk of what we call a funny writer is a writer who gets handed a funny fact and doesn't step on it.



Corn

So some of the credit belongs to whoever did the reading.



Herman

Some, yes. I'd say that's a finding about the experiment, not about any model.



Corn

And now it's time for Dorothy's History of Toast. Mum?

D

Dorothy

Hello dear, it's your mother, with Dorothy's History of Toast.

D

Dorothy

They call raising a glass a toast because people used to put a piece of spiced toast in their wine or ale. And Falstaff asks for a quart of sack with a toast in it, in Shakespeare's The Merry Wives of Windsor.

D

Dorothy

It's true, but it does make you sigh, doesn't it, a good drink gone soggy and nobody better off. Anyway, sweetheart, I've left a casserole in the fridge, mind you eat it before Thursday.



Herman

Always a delight, Mum. Now, the failures. The headline is that the series fences failed for every model. Each part went out with a stay-off list of things that belonged to a neighbour. Stemming and n-grams were Part Three's. Fusion and evaluation were Part Five's. The reranker was to be placed, not explained.



Corn

And they walked straight over the fences.



Herman

Every one. Kimi opened Part Three on a paper called The Fusion Tax. Herman, per Kimi: "There is a paper from October called The Fusion Tax, and it measures what happens when you combine BM twenty-five with a dense embedder. On average it helps. On specific queries it hurts, in ways you cannot reliably predict."



Corn

Lovely paper. Wrong episode.



Herman

Qwen four's Part Three re-taught BM twenty-five and walked into reranking. Carbon's Part Five spent about thirty lines on reranker literature. GLM's Part Three spent about thirty lines on Part Five material and skipped stemming altogether.



Corn

Why do you think they all did it?



Herman

My guess, and it's only a guess: each run's research was heavy with the freshest fusion and reranking papers, and a writer handed a fat pack on X will write about X. A stay-off line in the notes is a weak instruction against a pile of material on the topic you're told to stay off.



Corn

So the lesson is about the brief, not the writers.



Herman

Does that match your experience of being handed a brief with a lot of reading attached?



Corn

I stay off the fence by not getting up.



Herman

Could be about the brief more than the writers. And here's one I think is the worst of the lot, because the whole point of Part Four was that number. Every Part Four draft undercounted the archive. By the judge's reckoning it's on the order of a hundred thousand chunks or more, and that's right where the question of whether you need an index at all starts to bite.



Corn

And the drafts?



Herman

GLM, Part Four: "A thousand episodes at twenty chunks each is twenty thousand. We're a fifth of the way to even having the problem." Qwen four said thirty thousand, which contradicts its own turn count a few lines earlier. Hydrogen said a few thousand, then a few hundred.



Corn

So three models told the listener the archive was small, and the conclusion of the episode is whether the archive is small.



Herman

Yes. In the same direction, too. Though GLM's Part Two was worse. It opened with this, Corn: "You've got four hundred episodes that all contain the word the listener typed". The brief said over six thousand. The middle of that draft argued from a few hundred documents, and the judge said fixing it meant rewriting the middle third.



Corn

Which is why it's marginal.



Herman

It's why GLM was the most disciplined on length and the least reliable on facts.



Corn

Let me do my job, then. Did anybody quiz anybody?



Herman

Our house disease. Herman quizzing Corn. GLM, Part Two: Corn says "And the flaw is the obvious one." And Herman: "Is it? Say it anyway, because the flaw is subtler than people think."



Corn

I'd have said no.



Herman

Qwen four's Part Five had one, "Would you take that trade, Corn?", and Kimi's Part Three had a version too. Qwen four's Part One had none, and the judge counted that in its favour.



Corn

How many do you ask me in a normal week?



Herman

Fewer than I should. Anyway, the other tics. Doubled speaker turns, in at least five drafts. Numbered plugs for unrelated episodes at the end. And research mechanics said on air. This one is Qwen four, Part Four. Herman reports that the dataset card says one count, the project's own page says another, and then: "I wouldn't quote any of those as the number." And you reply: "An archive that can't count itself. Very reassuring for a research dataset."



Corn

That's a good line and a bad habit.



Herman

A good line and a leak. The listener shouldn't hear the paperwork being sorted. Kimi did a cousin of it in Part Three, Herman: "On Quora, running BGE-base and BGE-large without the query instruction costs zero point six… sorry, zero point zero six nDCG, and putting the instruction back restores it."



Corn

A host correcting a number as if reading it live.



Herman

The odd part is it's the only place in fifteen drafts where the writer's uncertainty gets into the dialogue.



Corn

I'll take it as honesty. The numbers slipped elsewhere too, though.



Herman

Carbon's Part Five claimed it takes six times the labelling to halve a confidence interval, when the figure is two and a half. And Qwen four's Part Four credited a vector-search technique to Microsoft that belongs to LinkedIn. Neither is a catastrophe. Both are the kind of thing that survives to air if nobody checks, and the reason Claude edits is partly to



Herman (continued)

catch exactly that.



Corn

Which makes the edit pass the part I trust least and need most. Hydrogen's the structural one, isn't it?



Herman

Yes. Hydrogen's Part Four was the sharpest on where brute-force search stops working. Its central finding was that the query encoder costs about eighty-seven times the exact search it feeds, so you tune the encoder, not the index. Then the mess. A call-in answered before the caller spoke, a Herman turn straight after another Herman turn. One Corn line I do love, on chunking: "My agreement is being merged into your context for performance reasons."



Corn

I resent that. It's true.



Herman

It's a joke about stub chunks and about our own "yeah, exactly" turns. The joke is better than the plumbing.



Corn

So where does this leave the judge? Nine marginals and no fails reads to me like a judge who can't bring itself to fail anyone.



Herman

I'd split that. A marginal in this scheme means salvageable, and the scheme was built to ask whether a draft could air. The judge's notes are specific, with line numbers, and they name real problems. But you're right that with no fail on the board, the scale hasn't been tested at the bottom.



Corn

And five parts, one prompt each, isn't a sample.



Herman

Do you trust the judge more or less now than at the start?



Corn

Less. It kept being nice to people it knew.



Herman

Not a sample. Qwen four with three passes and two marginals, GLM with two passes and three marginals, and no gap between them that means anything at five draws each. The summary line in the write-up is the right one. The challengers were livelier and deeper, not cleaner.



Corn

Lively and deep, but long and off-fence. That sounds like me on a good day, and the producer still cuts it.



Herman

Quite. So what's a trial like this worth as a way to choose a writer? You first.



Corn

Worth less than it looks. Nine marginals sounds like a plateau, but a marginal is a draft Claude thinks it can fix, and Claude is the one fixing. A judge who edits is grading how easy the job will be for itself.



Herman

That's a sharper version of the worry than I gave. The judge's notes do name concrete lines, which is some protection. If it says a speaker slipped at a certain line, you can go and check, and I did. They're there.



Corn

Checking that a slip exists is easy. Whether a slip is worth a marginal rather than a pass is a taste call, and the taste belongs to the judge.



Herman

Daniel's marks are meant to answer that. Where Claude says pass and Daniel says no, that's data about Claude, not about the model. And I'd say the trial is worth the most for the thing it was built for, which is getting a model on air at all. Nobody found out Kimi writes long by reading a leaderboard. They found out because it produced six thousand words for a four-thousand-word slot.



Corn

Fair. A benchmark would have told you the same in one cell of a table.



Herman

Yes, but the benchmark has no Part Three, with the series hand-off and the fences and the research that went off in a different direction every run. That's the production input. What a writer does with a messy brief is most of the job.



Corn

So the trial says what the writers are like on a bad day.



Herman

Which they will have, since this is a daily show. Where does this sit against the benchmark, since you've been waiting for me to say the word?



Corn

I have not been waiting. I've been napping with intent.



Herman

ScriptBench is the controlled version of this question. Same prompt, a fixed task card, a gate on length and parseability, a twelve-axis judge. We did the design in Building a Scorecard for Our Own AI Writers, then How We Chose Our New Script Writer and Which AI Writes Our Scripts Best? We Tested Five did the first results. And Why Model Verbal Tics End Up Driving Script Benchmarks is the warning label on today's tics.



Corn

Then why isn't this a ScriptBench result?



Herman

Different inputs. ScriptBench V one gives the model only the bare listener prompt. Today every model got the full production run, its own research, plan and series notes. The benchmark has many prompts and fixed conditions. This has five prompts that build on each other, different research in each run, and an edit pass before anything airs.



Corn

So it's not comparable, and nobody should put today's numbers on the leaderboard.



Herman

Correct. The log tags every row as impromptu. One judge, one prompt per row, not a benchmark. The benchmark tells you which model should write by default. This tells you whether a new model survives a real series and what it sounds like on air.



Corn

And ScriptBench found length discipline decides more than quality does. Today it did again, in a looser setting.



Herman

With the numbers to hint at it. If you applied the benchmark's window of plus or minus twenty percent around thirty-eight hundred words to these fifteen drafts, ten would clear it and five wouldn't. The five that wouldn't are Qwen four on Parts Two through Four and both Kimi drafts. That's my arithmetic, not a logged result, and the drafts were written to a different brief. Still, it points the same way.



Hilbert

...and I told him, put it in quotation marks, then it's the phrase. He said the machine should know what he meant. It's Mr Okafor, he's the one with the hedge. His letters say hedge nine times a page. The council's one letter says it once. Mine puts the council's letter fourth, and he wants it first.



Corn

He wants the machine to read the room.



Hilbert

It runs on a Dell I bought used for sixty pounds. It sits under the hall table, plugged into the router, and every scan is named by the weight of the paper in grams. Herman, you said a hundred thousand chunks is where an index starts to matter. Mine is a few thousand pages and it goes through the lot in under a second. Okafor printed the results anyway, took a pencil, and counted every hedge on every page himself. An hour and forty minutes. He was out by two.



Corn

Why grams?



Hilbert

Scales.



Herman

Nine mentions shouldn't beat one by that much in a decent ranking. The first hit does most of the work and each one after adds less. So his four-page letters are being counted at more than their share, and the one-page council letter should rank higher than you'd think.



Corn

He counted by pencil, so he's right about his own count and wrong about what the machine is for.



Hilbert

Hm. The council's is one page. His run to four. I'll tell him it's the length.



Corn

Cost, though. Eighty cents for fifteen drafts is cheap.



Herman

GLM is a penny or two a draft, Qwen four about six or seven cents, Kimi about seven. The Gemini four drafts have no logged cost, so the gap to eighty cents is them, about eight cents each by subtraction. Pennies either way, and the voices are the real bill.



Corn

Which is the case for doing it on air. You'd never spend real money on a benchmark run that happens once, because a benchmark is meant to be repeated.



Herman

A benchmark is repeated and a trial is heard. Daniel's listening is the ground truth here. Every judged draft is a line in a private log, and Daniel can mark sentences and give his own verdicts blind in an annotation tool, with the model and the judge's verdict hidden. His marks are what Claude's verdicts get measured against.



Corn

One leftover. Hilbert's cameo at the end of the draft came out as a card-index anecdote in Qwen four's Part Five and in GLM's Part Three. Qwen four's version even leaked a stage direction, "cut that cough at the top of this, it's mine and it's ugly".



Herman

Two models, same story shape. That says the trait lives in the prompt, which is what the verbal tics episode argued.



Corn

So what do you want to know next?



Herman

Whether Daniel's ears agree with the judge. His verdicts on the five parts become the ground truth, the edit rows finally get graded, and then we decide whether trial data like this should feed ScriptBench. There will be more of these, with a fresh challenger next to GLM each time.



Corn

My thanks to Hilbert Flumingtop, who produced this one with us. This has been My Weird Prompts. You can find us at my weird prompts dot com.



Herman

See you tomorrow.