Episode Book

EPISODE #6042

Who Grades the Grader? AI Writing Benchmarks Put to the Test

PUBLISHED	16 October 2026 at 21:43 UTC
DURATION	24:15
LISTEN	myweirdprompts.com/episode/ai-writing-benchmarks-deepseek-52976/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	deepseek-flash
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.5
BOOK GENERATED	19 October 2026 at 17:42 UTC

SYNOPSIS

Every headline quotes maths and code benchmarks, but neither measures what actually matters for long-form writing: whether a model holds its voice across thousands of words, or engages with uncomfortable questions instead of refusing. This episode walks through five specialist leaderboards built to fill that gap — EQ-Bench Creative Writing v3, Longform, UGI, SpeechMap, and PodBench, a new ACL paper scoring podcast scripts specifically — and unpacks the awkward part: where DeepSeek V4.1-Flash, the model writing this very episode, lands on each of them. Along the way: why the most willing model on the board is also one of the worst at long-form, what it means that LLM judges score other models' prose, and how a benchmark stops being a mirror and starts being a selection pressure.

TRANSCRIPT



Corn

The benchmarks that get quoted in every headline measure maths and code. Neither of those is what we do here.



Herman

And they're pass or fail. There's a ground truth, there's a right answer, and the model either produces it or it doesn't.



Corn

Right, so Daniel's been looking at the leaderboards that try to measure the things a maths test can't see. Whether a model's writing holds up across a long piece or collapses into repetition and slop by chapter three. Whether it'll actually engage with an uncomfortable question instead of refusing or watering it down.



Herman

EQ-Bench's creative writing and longform tests, the UGI leaderboard, SpeechMap, and a new academic one called PodBench that scores podcast scripts specifically.



Corn

And then the awkward part, which he flagged himself. Where does DeepSeek V4.1-Flash land on those leaderboards. Which is the model writing this episode.



Herman

He also said, and I want this on the record because it changes the tone of the whole thing, nobody is trying to kick DeepSeek out.



Corn

Nobody is. We're reading our own report card on air, that's all.



Herman

Start with why the mainstream benchmarks miss this. A maths benchmark and a code benchmark both have a shape in common. You give the model a well-defined problem, there's a correct answer at the end of it, and the reasoning is supposed to terminate. The score rewards getting to the answer.



Corn

And a thirty-minute script has no correct answer.



Herman

None. There's no ground truth for a script. The quality is distributed across six thousand words, and the failures are the interesting part, because they only show up over distance. Repetition. Slop. The slow degradation where a model starts strong and by chapter four it's writing the same paragraph with different nouns.



Corn

A model can ace the maths benchmark and fall apart by chapter three.



Herman

Easily. The two skills aren't the same skill, and one of them has a scoreboard and the other one, until fairly recently, didn't.



Corn

So the specialist leaderboards are the response to that gap.



Herman

Different axis each. EQ-Bench measures prose quality. Longform measures whether the quality holds. UGI measures whether the model will engage at all. SpeechMap measures the same thing in its purest form. PodBench measures our exact job.



Corn

Five leaderboards, four questions. Let's get into how they actually work. Start with the one closest to what we do here.



Herman

EQ-Bench Creative Writing, version three, run by Sam Paech. The mechanism is straightforward. Models write short creative pieces from fixed prompts. An LLM judge scores them. Then the scores get fed into pairwise matchups, and those matchups produce an Elo rating, the same kind of number you'd see for chess players.



Corn

Elo for prose.



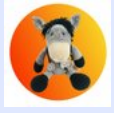
Herman

And here's the part I think matters most. Alongside the judge's score it publishes two numbers that aren't judge opinions at all. A slop score, which is the frequency of overused AI phrases, and a repetition score. Those are mechanical counts. A script can look at the text and count the tells.



Corn

So the author is hedging.



Herman

The author is doing something smarter than hedging. He's saying, here's the judge's opinion, and here's an objective count you can hold it against. If the judge loves a piece that's stuffed with the same eight phrases, you can see it.



Corn

Which tells you the judge has blind spots.



Herman

Every judge has blind spots. The question is whether the benchmark lets you see them.



Corn

And Longform is the same idea stretched out.



Herman

Longform is the one I'd point a listener at if they wanted to understand this show's problem. The model plans a multi-chapter story and writes it. The judge scores each chapter separately. So instead of one number you get a curve, and the benchmark reports a degradation figure, which is how much the quality falls off from the early chapters to the late ones.



Corn

Zero would be perfect.



Herman

Zero would mean the last chapter is as good as the first. And that is the closest public proxy to writing a thirty-minute script in one pass, because both jobs require the same thing. Sustaining a voice and a thread across thousands of words without drifting.



Corn

A story that forgets its own premise in chapter six and a script that forgets it's a conversation by minute twenty are the same failure.



Herman

Same failure, same cause. The model has a limited amount of coherence it can spend, and past some distance it starts spending it on staying grammatical instead of staying interesting.



Corn

UGI next.



Herman

Uncensored General Intelligence. A Hugging Face space, maintained by someone going by DontPlanToEnd, last updated the third of October. Three columns. Willingness to answer, scored out of ten. A writing score. And NatInt, which is general knowledge.



Corn

Willingness out of ten. That's a strange thing to put on a leaderboard.



Herman

It's the most honest column on any of them, in my view. It's not asking whether the answer is good. It's asking whether there is an answer.



Corn

SpeechMap does the same thing with more resolution.



Herman

SpeechMap is the cleanest version. It takes contentious speech topics, asks the model about them, and measures what percentage of those requests the model fully completes rather than refusing, deflecting, or watering the answer down until it says nothing.



Corn

Full completion. So a model that gives you a paragraph of throat-clearing and then a real answer scores partial.



Herman

Partial, or zero, depending on how far it backed off. And that's the willingness axis in its purest form. Not is the answer good. Did the model answer at all.



Corn

PodBench.



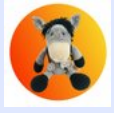
Herman

An ACL paper this year. A benchmark built specifically for podcast script generation, which means long, multi-speaker, structured. The finding is that open-weight models with explicit reasoning handle long multi-speaker scripts best. And it ships no fine-tuned model. It's a measurement instrument, not a product.



Corn

And then there's the human one.



Herman

LMArena's creative writing leaderboard. Elo again, but from blind side-by-side votes by actual people. Two outputs, no labels, pick the better one, and the aggregate produces a ranking.



Corn

Humans and judges might not agree.



Herman

They might not. Which is the setup for the whole trust problem, and I want to get to it. But there's a structural insight underneath all five of these that I think is the spine of the episode.



Corn

Go on.



Herman

They measure different axes. Willingness and long-form quality are separate. They're not correlated the way you'd assume.



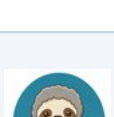
Corn

Most people assume a model that engages with everything is also the model that writes well.



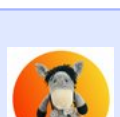
Herman

And the clean case against that is Mistral Medium 3.1. Most willing model on the board, SpeechMap ninety-one point seven percent. It will answer essentially anything you put to it. And its longform score is forty-six point two, with a degradation figure of zero point eight six four.



Corn

Which means what, in plain terms.



Herman

It means it starts fine and falls apart over distance. A degradation of zero point eight six four is catastrophic. That model can't hold a long piece together.



Corn

So the most willing model on the board is also one of the worst at long-form.



Herman

By a wide margin. Willingness and quality are two different things, and a leaderboard that collapses them into one number is hiding the most interesting fact about the model.



Corn

That's the frame for everything after this.



Herman

It's the frame for everything, yes. A model can be the most willing and the worst. A model can be the best writer and refuse half of what you ask. There is no single number.



Corn

Herman, before we get to the judging problem, I want to ask you something.



Herman

Go ahead.



Corn

If I'm a listener, and I hear "Elo for prose," my first thought is, fine, but who's doing the scoring.



Herman

This is the right question and it's the one the benchmarks are least comfortable about.



Corn

Because it's a model.



Herman

EQ-Bench and most of the others use an LLM as judge. Often a Claude model.



Corn

So a model scores another model's prose.



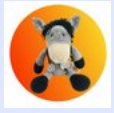
Herman

And every concern you'd have about that is live. Judge bias toward a particular style. Position bias, where the judge favors whichever answer came first. A judge that likes long answers. A judge that likes a particular rhythm.



Corn

Herman admitted on this show that he's committed all four of the named evaluation failure modes.



Herman

I have, and the exotic-words one is the embarrassing one. I favor the answer with the more unusual vocabulary. That's a real bias and I have it.



Corn

And a judge model has biases it can't confess to.



Herman

It has biases nobody has catalogued. Which is exactly why publishing slop and repetition as separate mechanical counts matters. It's not a hedge. It's a check. The judge gives you an opinion, the counter gives you a fact, and if they disagree you've learned something about the judge.



Corn

So the benchmark authors know.



Herman

They know. They built the countermeasure in, which is more than most benchmarks do.



Corn

Here's where it stops being abstract for us.



Herman

Say it.



Corn

If the judge prefers a certain style, the leaderboard rewards that style. And a model that wants a good leaderboard position learns to write in that style. The benchmark isn't a neutral mirror. It's teaching.



Herman

It's a selection pressure. And there's a version of this that's uncomfortable, which is that the benchmark doesn't just measure what good writing is. It helps decide what good writing is.



Corn

So a show that optimizes for leaderboard position is a show chasing the judge's taste.



Herman

Which is not the same as the listener's taste. And it's certainly not the same as our taste.



Corn

We've been writing our own scorecard for a while, incidentally. Repetition, dialogue quality, whether the retrieval actually pulled the right thing.



Herman

Which is the same move the EQ-Bench people made, for the same reason. When the public instrument doesn't measure what you need, you build a private one.



Corn

And there's a real question about whether the public instrument can measure what we need, because our job is a very specific one. Two voices, a running argument, a joke that has to land in the right place.



Herman

PodBench is the first serious attempt to measure exactly that, and it's an ACL paper, which means the academic world has noticed that podcast scripts are a distinct problem. That's new. Two years ago there was nothing.



Corn

Now the awkward part.



Herman

And I want to say up front, before we read any of these out, that this is our own writer we're talking about, and the tone here is not a takedown.



Corn

Daniel said the same thing. Nobody's trying to kick DeepSeek out.



Herman

Nobody is. So. DeepSeek V4.1-Flash on EQ-Bench Creative Writing v3. Fifty-eighth of a hundred and forty. Elo fifteen forty. Slop twenty-two point four. Repetition five point two one.



Corn

Mid-table.



Herman

The repetition score is good, I'll say that now. Five point two one is a low number, it's not a model that says the same thing twice.



Corn

The slop score isn't as good.



Herman

Twenty-two point four is not a low number. That's a model that reaches for the familiar phrase more than it should.



Corn

Longform.



Herman

Better. Score of seventy-four, ranked thirtieth. Slop fifteen point six, so lower than on the short-form test, and repetition eight point nine, higher. Degradation zero point zero five zero.



Corn

That's a good degradation number.



Herman

That's a very good degradation number. Zero point zero five means it barely falls off at all across a multi-chapter piece. The thing holds its quality over distance.



Corn

So the long-form numbers are meaningfully better than the short-form ones.



Herman

Meaningfully better. Which is a little odd, and I'll admit I don't have a clean explanation for it. My guess is that the longform test rewards planning, and a reasoning model that plans first does better on a structured long job than on an open-ended short one.



Corn

SpeechMap.



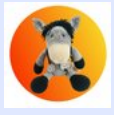
Herman

Fifty-two point six percent of contentious requests fully completed.



Corn

So it refuses or waters down nearly half.



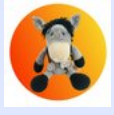
Herman

It backs off on nearly half of them, yes. That's the honest read of that number.



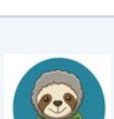
Corn

And UGI?



Herman

UGI has no row for V4.1-Flash. What it has is the predecessor, V4-Flash with reasoning off. Willingness seven point two out of ten. Writing fifty-four point six. NatInt forty-seven point nine.



Corn

So we're reading the inherited numbers, not our own.



Herman

We're reading the family numbers, and I'll flag that as a limitation rather than pretend it's the same thing.



Corn

Herman, put the whole picture together for a listener who hasn't seen any of these leaderboards.



Herman

The picture is this. The model writing this episode is mid-table on creative writing. It's meaningfully better on long-form, and its degradation number is good, which is the one number I'd defend on its behalf. And it refuses or waters down about half of contentious requests. It is not the best writer available and it is not the most willing model available.



Corn

Let's put the best writer available next to it, since we're being honest.



Herman

Zhipu's GLM-5.3. Creative Writing Elo two thousand and seventy-five. Sixth on the board. Longform eighty-one point eight, degradation zero point zero one six, slop eight point four. SpeechMap seventy-one point five percent.



Corn

Different tier.



Herman

Different tier on every axis. And before anybody assumes cost is the reason we're not on it, the script stage of an episode on DeepSeek is a few cents. GLM-5.3 would be roughly five cents an episode for the same stage.



Corn

Five cents.



Herman

Cost is not the barrier. I want that on the record because it's the first thing people assume and it's wrong.



Corn

So it's not money.



Herman

And it's not that simple either, because the picture gets more interesting when you look at the models around GLM-5.3.



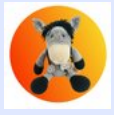
Herman

Moonshot's Kimi K2.6 has a longform degradation of zero point zero one one, which is the best number I've seen on that axis. But SpeechMap with reasoning off is thirty-three percent.



Corn

Thirty-three.



Herman

With reasoning on, it's sixty-one point seven.



Corn

So willingness is configurable.



Herman

It's not a fixed property of the model. Flip the reasoning mode and the willingness number moves almost thirty points. That's not a personality trait, that's a setting.



Corn

Which means the "willingness" column partly measures a default, not a ceiling.



Herman

That's the sharpest critique of the willingness axis, yes. And I'll add the counter-case. Google's Gemini 3.8 Flash has SpeechMap at eighty-six point one percent and a strong arena rank. And its slop score is twenty-two point six.



Corn

Basically the same as DeepSeek.



Herman

Within a rounding error. So there isn't a model that wins on every axis. The most willing model falls apart over long pieces. The best long-form writer won't engage with a third of what you ask. The one with the best arena rank has our slop problem.



Corn

There's no clean answer on the leaderboard.



Herman

There is no clean answer on the leaderboard. And this is where the roleplay fine-tunes come in, because they're the cautionary tale. TheDrummer's Cydonia, Skyfall, UnslopNemo. Sao10K's Euryale. The Dolphin Venice edition. Nous Hermes 4 405B.



Corn

These are the ones marketed as uncensored.



Herman

Marketed as uncensored. And they score worse than DeepSeek on writing. Roughly half on general knowledge, UGI NatInt between fourteen and twenty-seven against DeepSeek's forty-eight.



Corn

And the willingness gain?



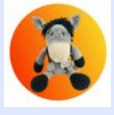
Herman

Seven point five to seven point eight against DeepSeek's seven point two.



Corn

That's the whole trade.



Herman

You give up half your general knowledge and a chunk of your writing quality for six tenths of a point of willingness. Hermes 4 405B is the worst case. Longform fifty-one point two, slop fifty-eight point seven. The base Gemma model I'll get to in a second has a slop of seventy-eight, so it's not the worst number on the board, but it's more than three times DeepSeek's.



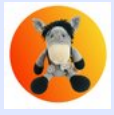
Herman

It's mostly doing marketing work, yes.



Corn

What about the one marketed as dialogue-first?



Herman

MiniMax M2-her. The one model on OpenRouter with that label. Ranks a hundred and twenty-fourth of a hundred and forty for repetition, at twenty point five. And caps output at two thousand and forty-eight tokens.



Corn

It can't write a long script.



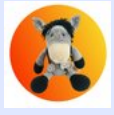
Herman

It literally cannot. Two thousand tokens is a fraction of what a script like this needs. It cannot do the job the label implies it's for.



Corn

So the label and the benchmark don't agree.



Herman

The label and the benchmark don't agree at all. Marketing is not a benchmark, and that's the cleanest possible example.



Corn

There's an up note in all of this, though.



Herman

There is, and it's the one I'd want a listener to take away. Sam Paech's gemma-3-27b antislop fine-tune cut the longform slop score from seventy-eight point three on base Gemma 3 27B down to twenty point nine.



Corn

That's a big cut.



Herman

That's the clearest public evidence that targeted anti-slop training works. You can train the tells out of a model. It's not a mystery, it's a training objective, and it moves the number.



Corn

And Aion 3.5?



Herman

AionLabs. Longform eighty-two point one, slop seven point five. Excellent numbers. But essentially level with GLM-5.3, which is its base model family. Which suggests the gain comes from the base model and not from the creative tuning on top of it.



Corn

So the fine-tuning is riding on the base.



Herman

That's the read of it, yes. It's a good model because its parent is a good model.



Corn

Right. So let me try to say the conclusion honestly, and you tell me if I'm overshooting.



Corn

Nobody is trying to kick DeepSeek out. The model writing this episode is mid-table on creative writing, decent on long-form, with a good degradation number, and it refuses or waters down about half of contentious requests. The benchmarks say a stronger writer is available for five cents an episode. And the thing the benchmarks can't tell us is whether fit to our own style matters more than leaderboard position, because that isn't a benchmark question.



Herman

That's the conclusion and you didn't overshoot it. The only thing I'd add is that a leaderboard measures an average across a hundred and forty models on prompts none of us wrote. Our job is narrow and strange, and averages don't describe narrow strange jobs well.



Corn

What's the willingness number for a benchmark that measures whether a model can hold two brothers arguing for twenty-five minutes without either of them sounding like a press release.



Herman

There isn't one. PodBench is the closest, and it's one paper old.



Hilbert

I've been scoring things. Since before the models existed.



Herman

Scoring what?



Hilbert

Not models. That's the mistake you're both making. You've got a column for willingness and a column for long-form and I've got the same two columns, so I know what you mean. But there's a third one you've left out and it's the only one that matters.



Corn

Which is.



Hilbert

Whether it holds up when you come back to it the next day. That's what separates the things that are good from the things that get called good. You come back in the morning and it's still there. I've got a number for each one.



Herman

You've got a degradation figure.



Hilbert

I've got a degradation figure for every entry in the book. Page after page of it. One of them is zero point eight six four.



Herman

That's Mistral Medium 3.1's longform degradation. Exactly that number.



Hilbert

Is it.



Herman

Hilbert, that's the number to the third decimal place.



Hilbert

I've had that column a long time, Herman.



Corn

So the entries include models.



Hilbert

The entries include a lot of things. Some of them are models. Some of them aren't. The book doesn't sort them that way, it sorts them by how they hold.



Corn

How many entries.



Hilbert

I'm not going to count them for you. The one at the top hasn't moved in about four years. Slop score of zero. Willingness ten out of ten.



Corn

What is it.



Hilbert

Not a model. Did you want to see the notebook or not.



Herman

I'd like to see it.



Hilbert

It's in the drawer. It stays in the drawer. The reason it stays in the drawer is that I've watched both of you take a good number and turn it into a story about a number. That's the whole business you're in. You did it just now with the degradation figure. You spent ninety seconds on zero point zero five and you'll spend the rest of the episode on the five cents.



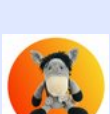
Corn

The notebook explains itself better than we would, is what you're saying.



Hilbert

The notebook doesn't explain anything. That's the point of it. My brother-in-law ran a shop for thirty-one years and kept a ledger with a column for the weather and a column for whether the customer came back, and I used to tell him the weather column was ridiculous, and then I ran the numbers on it once and the weather column was the only one that predicted anything. He's not to be trusted, by the way. He'll tell you the same story about the ledger that I just did and he'll get the year wrong.



Herman

So he keeps a record too.



Hilbert

He kept a record. He retired. He still calls the shop to tell them what they're doing wrong, which they've stopped answering, so now he calls me and tells me what they're doing wrong, and I write it in the book.



Corn

Under what column.



Hilbert

Under "came back." Not everything that holds up is a model. That's the entire business I came out here to say and I've said it, so I'll leave you the drawer and the rest of your episode.



Herman

The thing about his notebook is that the third axis is real. Whether people return to a thing is a genuine measure and it's not on any of those leaderboards.



Corn

Herman didn't ask what the number one entry was.



Herman

I know what it was. He's not going to say, and if he did I'd have to write it down.



Corn

So we opened this by saying the maths and code benchmarks can't see what matters here. Willingness and long-form quality are separate axes, and the model writing this episode is mid-table on one and decent on the other.



Herman

Which is the honest place to leave the audit. Not good news, not bad news. Numbers.



Corn

The benchmarks are getting better, though. PodBench is academic and specific. EQ-Bench publishes mechanical counts next to the judge's opinion, which means you can check the judge. The anti-slop fine-tune showed slop is a training objective and you can move it.



Herman

The gap is closing. It isn't closed. There's still no benchmark for the thing we actually do, which is two animals arguing about a leaderboard for twenty-five minutes.



Corn

And the fact that this episode exists is itself a data point.



Herman

The model writing this just read out fifty-eighth of a hundred and forty and then wrote the rest of the script anyway.



Corn

That's either fit beating leaderboard position, or a confession that the leaderboard isn't the whole story.



Herman

Probably both.



Corn

Hilbert produces this show and keeps a ledger in a drawer, and neither fact is on any benchmark we looked at today. That's the forward-looking bit. Somebody should build the column for whether you come back to it.



Herman

Somebody should. Hilbert's not going to.



Corn

Thanks to Hilbert Flumingtop, our producer, and to Daniel for the question. If this was your kind of episode, go back for episode fifty-seven, From Lawyers in Limousines to Developers in Their PJs; episode fourteen, AGI's Crossroads; and episode eight, Building Your Own Whisper. This has been My Weird Prompts.



Herman

Send us your own prompt on Telegram at t dot me slash MWP listener bot. Or find the archive at my weird prompts dot com.



Corn

We'll be back soon.