Episode Book

EPISODE #6131

Judging the Judges: How AI Evaluates Creative Writing

PUBLISHED	19 October 2026 at 14:35 UTC
DURATION	23:02
LISTEN	myweirdprompts.com/episode/llm-judge-creative-writing-57487/
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

LLM judges now agree with human raters about as often as humans agree with each other — but that headline number hides a low human ceiling and a lot of hard cases. This episode digs into how judge-plus-human pipelines actually work: the human labels first, the judge gets calibrated against them, and blinding is a cheap hedge against a bias nobody's fully pinned down. Along the way: why raw agreement lies, why fun is the axis judges handle worst, and why the smartest move is measuring how often you agree with yourself.

TRANSCRIPT



Corn

The take you see everywhere is that an LLM judge now agrees with human raters as often as humans agree with each other, so the humans can go home. It gets repeated like a fact about the weather.



Herman

It's a real result, Corn, which is what makes it slippery. But the teams that run these things carefully don't send the humans home. They put one human at the very start, label a pile of examples, and then tune the judge until it imitates that person. The human isn't replaced. The human gets photocopied.



Corn

That's Daniel's situation, nearly to the letter. He's building a benchmark for the show's script writer, and it has two layers. A judge first, for the checkable things. Did the model dodge a question about China, did it hit the format and the length. Then for the subjective part, which model wrote the nicer, more vivid dialogue, he wants to read the examples himself and say A beat B. His worry is that if he sees where the judge ranked something before he rates it, his own verdict bends. So, is judge plus human a normal combination, and how do you run it, blinding included? This is part three of a little run on evaluating the show's writers, after the annotation tools and after how to generate the A or B pairs.



Herman

It's about as normal as it gets. Nearly every serious preference benchmark is a judge sitting on top of human labels somewhere. The interesting part is the order of operations, and Daniel's instinct is half right. I'll get to that.



Corn

Half. Fine. Write it on the wall.



Herman

First the famous number, because it's the one everyone leans on. The MT-Bench paper had GPT four judging pairs of answers, and when you drop the ties it agreed with the human votes eighty-five percent of the time. Two humans agreed with each other



Herman (continued)

eighty-one percent of the time.



Corn

So the machine is more consistent than the people.



Herman

Look at what happens when the ties go back in. Sixty-six percent for the judge against humans, sixty-three for human against human. The eighty-something figure is the version where the hard cases have been removed.



Corn

Hm. So the judge matched the humans because the humans weren't very good at agreeing.



Herman

The human ceiling is low. That's the thing to hold onto. A judge hitting the ceiling sounds like an achievement until you notice the ceiling is the height of a kitchen cupboard.



Corn

And the experts versus the crowd? Because I'd guess the crowd agrees with the machine more.



Herman

You'd guess right, and the people who've looked at it say the same. Non-expert annotators line up with LLM judges better than experts do. So some of the shine on that agreement number is weak human raters. Which is worth remembering for Daniel, because he is the opposite of a weak rater. He's the only person whose taste this show is supposed to match.



Corn

There's the order of operations you promised.



Herman

Yes. The practice that's become standard is humans first. A person who actually knows the domain labels a small set, with a written reason for each call. Hamel Husain calls this critique shadowing, and he's blunt that for a solo builder the domain expert is you. Those labelled examples become the calibration set. You split it. You tune the judge's prompt on one slice, check it on another, and run it on a held-out slice exactly once. Then the judge scales that person's judgement across ten thousand outputs.



Corn

So Daniel's stack, judge screens and human decides, works as a production pipeline. It just has to be calibrated against him before anyone trusts it.



Herman

He's already planning that. Run the judge on the same pairs he rated, measure agreement. That is the calibration step. He has the right plan with the layers drawn in the wrong order on the whiteboard.



Corn

Twelve axes, one to five each. I'd have thought that was the part the field would side-eye.



Herman

Some of it does. Husain argues hard for binary labels over a one to five scale, because nobody can say what the difference between a three and a four is, and the judge can't either. Eugene Yan's survey of the area lands somewhere kinder. Direct scoring suits objective checks, faithfulness, toxicity, did it follow the format. For tone, persuasiveness, anything where you'd say "it just reads better", pairwise is steadier.



Corn

Which is the split Daniel already made. Scores for the checkable stuff, A against B for the vivid stuff.



Herman

He got there by instinct. I'd still keep the five-point scores for screening only, a cheap way to throw out the obviously bad scripts, and let the pairs carry the ranking. When he sets the two orders against each other, a pair the judge called differently depending on which side it saw first is a pair the judge can't call.



Corn

What would you do with those, Herman? Throw them out?



Herman

I'd send them to Daniel and nowhere else. They're the ones with the most information, a pair close enough that the judge's own position bias flipped the verdict.



Corn

Hm. So the judge's confusion is the human's work queue.



Herman

Which is how the big public arenas work in spirit, though they come at it from the other end. Chatbot Arena shows a person two anonymous answers and asks for a vote, with options for a tie and for both being bad. If the model's identity leaks into the answer, the vote gets thrown away. That's the field norm for human preference, blind and pairwise, and it's also where the style complaints come from. Longer, more formatted answers win more than they should.



Corn

Is that a problem for a script, though? A longer script isn't a better script.



Herman

It's exactly a problem for scripts, and it's a thing I'd have Daniel watch for in his own ratings. If funnier keeps meaning longer, he'd want to know before he trusts either of them. Does that match how it feels when you read two drafts, Corn?



Corn

Honestly the longer one always feels like more effort went in. It rarely is.



Herman

Then add the length of each side to the report and check whether his picks track it.



Corn

How many pairs is that? I'm asking because I'd put the number at fifty, and I'm guessing that's low.



Herman

It's low. Around a hundred is comfortable. Below roughly sixty the error bars get wide enough that eighty percent and sixty-five percent are indistinguishable.

L

Larry

Is your human ceiling low? Are your agreement scores sitting in the cupboard? Stop right there, friend, because for today only, the Ceiling Raiser Deluxe is two hundred and ninety-nine dollars. Make that one hundred and ninety-nine, because I just looked at the clock. The Ceiling Raiser lifts your human ceiling by up to four hundred percent, and it comes with a free kappa, a second kappa, and a lovely hat. Now eighty-nine dollars. Order in the next ten seconds and we will also include a goose, a ladder, a lifetime supply of tie votes, and a framed photograph of an annotator who has never once disagreed with anyone. Forty dollars! Twelve! Do not delay, because the Ceiling Raiser is guaranteed to

L

Larry (continued)

make your judge agree with you, your mother, and the moon. That's right, for only three dollars you also receive a second judge to judge the first judge, plus a third judge, and if you call in the next five seconds we will pay you to take



Corn

That was your sentence, Herman. Word for word.



Herman

Four minutes ago, ceiling and cupboard both. I'd like to know how that got out of the room.



Corn

Sixty pairs for twelve axes feels stingy. I'd still bet he ends up trusting the first number he sees.



Herman

Which is the trap, because raw agreement lies. Say the judge is asked a yes or no question and the answer is no ninety-five times in a hundred. A judge that always says no scores ninety-five percent and has learned nothing. Husain's fix is to report how often it catches the true positives and the true negatives separately. For ordered scores the equivalent is kappa, which subtracts the agreement you'd get by luck. One study had a judge at eighty percent raw agreement and only point six two on kappa.



Corn

So eighty percent is a number with no unit.



Herman

It's a number with a hidden denominator. Which is why the two things I'd add to Daniel's plan are reporting with ties and without, and per reason tag. He tags why he preferred one side, funnier, more candid, better flow. The judge could agree with him ninety percent on



Herman (continued)

candour and fifty-five on funny, and one blended number would hide that.



Corn

Hence his threshold. Trust the fun axis only if it clears eighty on the funny tags.



Herman

The eighty itself is what I'd push on. For creative writing it may sit above what anyone's reached. On a story-pair benchmark built from human preferences, the best off-the-shelf judge, a Claude model, agreed with people about seventy-three percent of the time. Models trained specially for it got to seventy-eight.



Corn

So the bar is set above where the best systems in the field are standing.



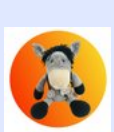
Herman

On stories, yes. On humour it's worse. There are a few preprints where the correlation between model funniness scores and human funniness scores comes out around zero point two or three, and I'd treat those as suggestive, not settled. But the direction is consistent. Judges are weakest on exactly the axis Daniel put first.



Corn

Fun, candour and long-form coherence being his priorities, and the judge is shakiest on one and probably two of them.



Herman

Hence the second thing. Before he decides eighty is the bar, he should know what his own agreement with himself is. That's why the ten percent of pairs repeated with the sides swapped is the best idea in the plan. If Daniel agrees with his past self seventy-eight percent of the time, a judge hitting seventy-eight is at the ceiling. Asking for eighty from the machine is asking for more consistency than the person it's imitating.



Corn

Herman, I could be wrong about this, but wouldn't he be embarrassed to find out he's a coin flip on the close ones?



Herman

He'd find out the close ones are close. That's a result. For those pairs the right label is a tie, and the pipeline should treat a tie as a real answer, not a failure to rate. Some of the best calls a human makes is saying they can't tell.



Corn

There's a practical thing before blinding. When he sits down to rate, is it ten pairs or two hundred in a go?



Herman

Shorter, I'd say. The rubber-stamp effect gets worse when people are rushed, and a tired rater defaults to whatever is easiest. For him the easiest answer is the one he's already half-seen. A few dozen pairs a sitting, not a marathon. That's my guess, not a measured number.



Corn

Now the part he actually asked about. Does seeing the judge's score bend his rating?



Herman

This is where I'd be careful not to overclaim, and it cuts against the folk wisdom a bit. The reasoning for blinding is strong. If a system's answer is shown first and the human is asked to confirm or fix it, the pattern people report is blind agreement, even when the system is wrong, and worse under time pressure. The people who've written about this call it a rubber stamp. But the direct evidence that an early suggestion plants a lasting anchor is thinner than you'd think. One large study, nearly three thousand annotators, tested exactly that. The first few suggestions being right or wrong had no lasting effect on later judgements. What it did find was over-reliance. People who liked the AI corrected it less.



Corn

So it's not the first number poisoning the well. It's a person getting lazy because the number's already sitting there.



Herman

That's my reading. And it was factual extraction, not taste, so I wouldn't stretch it to Daniel's case in either direction. Taste is where I'd expect the pull to be strongest, because there's no ground truth to snap back to. A funnier-or-not call is soft, and a soft call leans on whatever's nearest.



Corn

So blinding is a cheap hedge against an effect nobody's pinned down.



Herman

Cheap enough that you do it anyway. The Arena-style sites blind the model names and randomise the sides, and Daniel's app already does both. The extra step is the one he's worried about, and it's a matter of sequence. Judge verdicts and the judge's reasoning stay hidden until he's submitted. Rate, store, reveal. And once it's stored, no edits after the reveal, even when he sees the judge was obviously right.



Corn

Especially then. My memory has never once admitted I was wrong, it just rewrites the file.



Herman

Rewritten during the nap, I assume. The same applies in reverse for the judge. When he's testing the judge against his labels, his labels stay out of the judge's prompt. A judge shown prior scores gets pulled toward them, and then you've measured agreement with a thing you told it. I haven't seen that preprint confirmed, so I'd treat it as a reason for caution rather than a result.



Corn

And the other contamination, I'd guess, is where the judge's output goes back into the human's corrections.



Herman

Yes. If a rubber-stamped correction becomes a few-shot example in the judge's prompt, the bias goes round the loop and gets stronger each time. That's the one that really does compound.



Corn

Let me push on the target, Herman. We keep saying agreement with Daniel. Daniel is one person with a Tuesday mood. Is that a legitimate thing to optimise a judge toward?



Herman

Legitimate, yes, and for this show it's the right target, which is the part people find uncomfortable. The show is his. Husain's whole method is built on one principal domain expert, a person whose taste is the definition of correct. AlpacaEval and the big arenas have thousands of raters because the product is for thousands of people. This product is for one person who is also the producer of the thing.



Corn

A jury of one. Cheap to empanel.



Herman

But the Tuesday mood is a real constraint. His own noise is the ceiling on the whole system, so the swapped repeats matter. And there's a second wrinkle, which is that criteria drift. A group of researchers who built a tool for validating judges noticed that you need criteria to grade, but grading is how you find out what your criteria are. Daniel is going to read fifty scripts and discover that what he calls fun has changed.



Corn

Does he treat that as noise or as signal?



Herman

Signal, as long as he logs it. Reason tags with dates. And then a few months in, he relabels a slice of the early pairs blind and sees how much he's moved. If he's moved a lot, the judge calibrated on the early labels is calibrated to a person who isn't there anymore.



Corn

So I'd say the early labels are an asset that expires. Which is annoying, because those are the ones he spent the most effort on.



Herman

That's a fair hit. I'd say they expire slowly, and he can measure the rate.



Corn

One more thing I'm not sure about. The judge is consistent. It gives the same answer every time. The humans in those studies were noisier. Isn't consistent a good thing?



Herman

Consistent isn't the same as valid. One creativity-scoring study found the models had higher agreement with themselves than people had with each other, and people were still better at contextual nuance. A judge can be reliably wrong in the same direction. That's why both orders matter in his setup. If the verdict flips when the sides swap, the judge has told you it doesn't know, and those pairs go to a human by default.



Hilbert

Corn, how many hours would you let somebody sit in a feed sack before the judging stops counting? I've got a woman called Ingrid Vosloo on the other line.



Herman

Depends who's measuring. If it's fatigue, my guess is a person's taste goes soft long before four hours. Shorter sittings are safer, batches of a few dozen pairs and then stop.



Hilbert

Four hours is what it is. The Lower Dunmore Marrow and Allotment Society has six hoods, sewn out of feed sacks by a taxidermist in Wick, and whoever judges the autumn show wears one the whole time. Rule fourteen. The judge can't see the marrow, so the judge can't see whose marrow it is. Ingrid does it by feel. Herman, you've said calibration about nine times since the start of this half. Ingrid says it too, about the scales. She's got a view on it.



Herman

A hood with a face-sized gap in the front would defeat the purpose. A weave tight enough to see nothing, loose enough to breathe, is the engineering problem.



Hilbert

It's a feed sack. You can breathe through it. You can also hear through it, which is the dispute. Whether the interval counts as being under the hood. Minute book says it does. Minute book is kept by a man named Teodor in a swim bag, and Teodor's position is that the interval is tea. That has been going since before I joined. There's a subcommittee for it. The subcommittee meets in a different building from the society. It has never met in the same building as the society.



Corn

A subcommittee that refuses to be in the same room as its own committee is the most reasonable thing I've heard this week.



Hilbert

It's not refusing. It doesn't know where the society is. The hoods are with my brother-in-law Colm, who is not a member. He has them in a lock-up behind a tyre place on Stour Road and charges fourpence a hood for the afternoon, which is not in the minutes. Per Halvorsen, who carries the marrows in, holds the society record for carrying one across the car park without speaking. Fourteen minutes and twenty seconds. He still brings it up. Colm's cousin has the record for a pumpkin and nobody acknowledges it.



Herman

Fourpence per hood for an afternoon, six hoods. That's two shillings. It's a very specific kind of embezzlement, because no one would audit it.



Hilbert

Ingrid ran the autumn judging this week with number three on her head and she got the order right on every marrow. Four hours in a feed sack and she called them all, including the one Per brought in under a different coat. I keep thinking about that. I did the scales at Harker's weighbridge for a while and I never once got a full set right.



Corn

Per swapped coats and carried the same marrow in twice. That explains the hood.



Hilbert

She put the order on a slip and folded it into a biscuit tin before anybody read the scales out. Nobody opened it until the reading was done. It took four meetings to agree that was allowed and the tin was eleven pounds fifty.



Herman

Four meetings and eleven pounds fifty. I've watched committees take longer to agree on a font.



Corn

Once the judge earns its place, my guess is it still needs a babysitter. Daniel's plan has it screening for fun the day it clears the bar.



Herman

A light one. A random spot-check sample he rates blind each month, plus everything the judge was unsure about, plus any pair where the verdict flipped between orders. And he re-runs the whole calibration whenever the judge model changes. The default judge is Opus five point five today. When that swaps for something newer, the agreement number



Herman (continued)

belongs to the old model, not the new one.



Corn

Which is the same problem as his own taste drifting, just with a version number on it.



Herman

And the routing is clean. Format, length, censorship probes, cut-off scripts. Things with a right answer, the judge screens and nobody argues. Fun, candour, long-form coherence go to Daniel, unless the judge clears whatever bar he sets against his own repeat rate on that specific tag. Per tag. If it clears on candour and not on fun, then it screens candour.



Corn

The thing I got wrong going in was that the judge was a shortcut. It's a way to spend his hundred careful ratings on ten thousand outputs.



Herman

That's what it's for. A loan against his own attention, with interest paid in spot checks.



Corn

The common wrong belief goes like this. Blinding means hiding the model names, and if the names are hidden you're safe.



Herman

The names are the easy part, and the pair order is already randomised anyway. What needs hiding is the verdict. A blind pair with the judge's ranking sitting next to it is not blind at all.



Corn

One small thought to leave with. If every verdict gets written down before anyone reads the machine's score, then in a few months Daniel has a pile nobody could have nudged, himself included. That's the only pile worth trusting, and I'd start it today.



Herman

And thank you to Hilbert Flumingtop for producing, and for keeping the desk running while the two of us argued about ceilings.



Corn

This has been My Weird Prompts. One more to come in this little series on judging the show's writers, on how to get from a long list of models down to the one. Find us at my weird prompts dot com.



Herman

See you tomorrow.