Episode Book

EPISODE #6133

One Rater Is Enough: Designing Daniel's Taste Test

PUBLISHED	19 October 2026 at 14:47 UTC
DURATION	23:13
LISTEN	myweirdprompts.com/episode/single-rater-taste-evaluation-60761/
HOSTS	Herman, Corn, 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

Most evaluation advice says to average away individual taste. For a show with one creator, that advice is backwards. This episode works through how to build an evaluation system where the authority is a single domain expert — and the reliability that matters is whether that expert agrees with himself three weeks later. Drawing on applied LLM evaluation practice, qualitative coding methods, and a creative writing study that tested expert agreement on short stories, we lay out a concrete design: open coding before criteria, versioned rubrics, binary labels instead of five-point scales, anchor scripts for calibration, and a side-by-side verdict that comes before the marks. Along the way we argue about whether a show that lives in audio can be graded on a page, and land on the honest answer: read to screen, listen to decide.

TRANSCRIPT



Herman

The standard advice on judging creative output is almost always the same. One person's taste is noise. Get more raters, average them, and the noise washes out. If you can't afford raters, build a metric.



Corn

And for a show with one creator, that advice is backwards. Averaging away the one opinion that defines the show is like polling the neighbours on how you take your tea.



Herman

Which is the actual problem Daniel's brought. Even with no budget ceiling at all, he can have every model on the market write a script for the same prompt, and he's still the bottleneck. Two forty-minute scripts side by side, and somewhere around the twelfth minute of the second one he can no longer say which is better, only that one felt more alive.



Corn

What he's judging is whether the dialogue is good, whether it's funny, whether it repeats. He can invent a heuristic for each, and he told us he doesn't believe any of them will reach his judgement, least of all as the person who made the show. So the question is how you design an evaluation that someone else could rerun, when the ground truth is one person's taste, and what the screen in front of him should look like while he does it. Call this a follow-up to the run of episodes on grading our own writers, because it picks up exactly where that run stopped.



Herman

The first thing to say is that his instinct is the recommended starting position, not a confession of weakness. The people who've done the most applied work on evaluating language model products, Hamel Husain and Shreya Shankar, say plainly: for anything taste-dependent, pick one domain expert and make that person the authority. They call it the benevolent dictator. One consistent signal, no annotator arguments, and you only bring in several raters when you're a big organisation that needs to measure agreement between them.



Corn

So I've been the sloth who needed a committee this whole time.



Herman

You've been the sloth who needed a nap. Different thing.



Corn

Fair.



Herman

But here's what changes. If the expert is the authority, then the thing you need to be able to repeat isn't agreement with other people. It's agreement with himself. If Daniel rates a script on a Tuesday and rates the same script three weeks later, does he give the same verdict? That's the reliability that matters here, and it has a proper name in the measurement literature, stability. A coder's ability to repeat their own coding of the same material.



Corn

Which is a more flattering bar than agreeing with strangers.



Herman

It's a harder bar in one way. Strangers can disagree with you and you can call them wrong. You can't call yourself wrong without it meaning your rubric is vague.



Corn

Before we get to the machinery, you said he can't just sit down and write the rubric. Why not? He's got twelve axes already.



Herman

He's got twelve axes because of something he did by hand. In October twenty-two Opus five point five episodes, written offline, aired next to twenty-two from the pipeline running DeepSeek V four point one Flash. Listeners found the Opus ones far more enjoyable and so did Daniel. The obvious guess was repetition. The close read said no. The gap was mostly in what the turns do. In the Opus episodes both hosts hold opinions, where in the other set one is a lecturer and the other is a cue-giver. Callbacks change meaning when they come back later. And scenes, instead of inventories.



Corn

Those became the axes and the deterministic markers, and the judge scores them from one to five. I remember that bit, it's the scorecard episode.



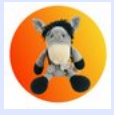
Herman

Right, and look at how he found them. He read forty-four episodes with a pen, wrote down what he noticed, and grouped the notes. That is the method in miniature. In the evaluation literature it's called open coding, borrowed from qualitative social science, and it's the first step in Husain's process. You journal on real outputs, you write down the first thing that's wrong in each one, and only afterwards do you cluster the notes into named categories and count them.



Corn

Why the first thing that's wrong? Why not everything?



Herman

Because the first problem is the one that would have made you stop reading. Everything after it is damage you only noticed because you were still being polite. And it stops a script that has six small problems from outweighing one that has a single fatal one.



Corn

Hm. That's closer to how I listen than how I'd grade.



Herman

The point I'd stress is that he can't write the criteria first. A paper from the user interface community, on tools that help people validate language model judges, called it criteria drift. You need criteria to grade, but grading is what teaches you what the criteria are. Their participants kept changing their minds about what they wanted after seeing outputs. Their conclusion was that you can't fully decide the criteria before the human judging begins.



Corn

So the rubric is never done.



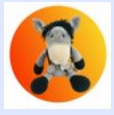
Herman

The rubric is done at a version number. That part is our addition, the paper doesn't talk about versioning, but it follows. After each round of coding you freeze what you've got as, say, version one of a taste document. Each named quality gets a one-sentence definition, one quoted example of it done well and one done badly. When round two changes a definition, you've got version two and you know that scores from before aren't directly comparable.



Corn

Right. I'll tell you where I disagree with where you're headed, because I can see it coming. You're about to say the twelve axes should be binary, and the per-script verdict should be spans of text, someone with a highlighter.



Herman

I was going to say binary, yes. Practitioners argue against one-to-five scales for this kind of work. Adjacent points mean different things on different days, people drift toward the middle, and you need more samples to see anything. A pass or fail on a named quality, with a one-line reason when it fails, is something he can actually repeat. The grade comes from counting how many of the named qualities pass. That's a practitioner opinion rather than something anyone has measured, I'd want that said.



Corn

And I don't buy that it should be spans. The thing he told us is that he can't pinpoint why one script is better. It's a gestalt. If I'm made to mark forty-one sentences with a highlighter, I've turned a gut verdict into homework and the gut is the thing he trusts.



Herman

The translation people ran into exactly this. The gold standard there marks error spans, with a category and a severity, and it's slow. So they trimmed it to a version where you mark the spans and then give one overall score, and it came out faster and cheaper at the same quality. The marks make the final verdict harder to fudge. You can't say the script felt flat without pointing at where.



Corn

Pointing at where is a skill I'd rather not exercise on a Sunday.



Herman

You wouldn't mark everything. You'd mark the first place you'd have switched off, and the best moment. Two marks. And then the gut verdict, in the form the same research found works best: a side-by-side call. Much better, better, the same. Side by side got more consistent answers than rating each item alone, every time they compared them.



Corn

Two marks I can live with. I'll hold the position, though, that the verdict comes first and the marks come after, because if I mark first I'm justifying instead of judging.



Herman

Then do that. Verdict, then marks, and the order gets recorded so we can see whether it matters.



Corn

You've just agreed with me and kept your own design.



Herman

It's a talent.



Corn

There's a number you keep dancing around. How noisy are people on this kind of thing, even experts?



Herman

A creative writing study from a couple of years ago had experts apply fourteen yes-or-no tests to short stories, three experts per story. Agreement between them on individual tests averaged what statisticians call moderate, about point four. The worst was a test called scene versus exposition, which is practically our scenes-not-inventories axis, and that sat at about point three. But when they added up how many tests each story passed, the experts' totals correlated well, around point seven.



Corn

So each single yes or no is shaky, and the pile is solid.



Herman

That's the lesson. Binary labels are noisy one at a time and stable in aggregate. For Daniel that's reassuring, since the aggregate is what ranks models. And in the same study the language model judges correlated with the experts at close to zero.



Corn

Which tells you the judge needs his labels, and that was the last series, so I won't make you do it again.

L

Larry

Friends, are you tired of dialogue that does nothing? Let me tell you about a satisfied customer. Gwendolyn from Ohio was reading her scripts in the dark, and then she found the Gut Verdict Pack. Gwendolyn said, and I'm reading from her letter, "I used to give everything a three, and now I give things a binary, and my marriage is stronger." Wonderful woman. Anyway, Brenda had never tracked a callback before the Pack, and now Brenda tracks a callback in her sleep. Brenda's husband Stanley says it changed his life, or somebody's life. Stanley, Brenda, Gwendolyn, whoever. The Gut Verdict Pack. Twelve axes, zero axes, results may vary. Ask your doctor whether your doctor can read.



Corn

"Ask your doctor whether your doctor can read."



Herman

Please don't do that again, Corn. It was worse the second time and it was slower.



Corn

Sorry. Back to the anchors.



Herman

Anchors, yes. Nobody calibrates a taste test on the day. Descriptive panels in food science train people against physical reference standards, a vial of something that is exactly the quality being named. In sensory research, with calibrated standards, two panels of eleven people tasting white wine produced profiles you could swap between them. In audio listening tests they go further. There's an international standard where listeners score clips against a hidden reference, plus deliberately degraded clips as low anchors, so that one lab's seven out of ten means the same as another's.



Corn

So he needs a shelf of scripts.



Herman

A gold set. Two or three episodes he loves, two or three he can't stand, and one deliberately bad one, a low anchor. Something built to have the summary-halfway-through problem, say, or both hosts agreeing on everything. Slip them into sessions blind. If the loved ones start scoring lower and the low anchor starts scoring higher, his scale has moved, not the writers.



Corn

And when the rubric changes version, they're the first thing he re-rates.



Herman

That's the move, yes. Re-grade the anchors, re-grade a sample, and only then compare numbers across versions.



Corn

Here's what I'd push on. Everything you've described is reading. The show is audio. Jokes live in timing, and a line that reads dry can play flat. If the product is listened to, shouldn't the grading be?



Herman

Listening is slow. Twenty-five minutes of audio is twenty-five minutes. He reads a script in maybe eight, and he can mark a span in a second. If every candidate had to be listened to, he'd get through a handful a week and the whole thing would collapse under its own cost.



Corn

But it's cheap to read a joke and love it, and hear it die in the voice.



Herman

That's a fair hit. Chatterbox has its own rhythm, and I'm not sure what it does to a written quip that depends on a beat. My view is you read to screen and listen to decide. Reading mode is for the broad sweep, listening mode for the finalists.



Corn

I'd move the line. I'd listen to anything he cares about winning on funny.



Herman

We might both be right. Funny is the axis I'd least trust to text, I'll give you that. But everything else on the list, the opinions, the callbacks, the scenes, you can see on a page.



Corn

Noted. Unresolved.



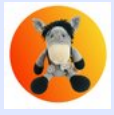
Herman

Unresolved. Which is why the screen needs both modes. Let me say what I'd put on it, concretely, because this is the part he asked for.



Corn

Describe the reading one first. I want to know how many buttons I'm going to resent.



Herman

The reading screen shows one script, nothing else. No model name, no date, no pipeline label, an item number that means nothing. The text takes the middle of the screen. He selects a span with the mouse, or moves a cursor by sentence with the arrow keys, and presses a number key. Each key from one to nine is a tag, a named quality from the taste document, and the chip shows up in the margin. Tags are the named failures and the named virtues, so one key might be lecturer turn, another is inventory instead of scene, another is callback that lands.



Corn

What happens on a fail?



Herman

A small text box opens, one line, optional but nudged. Why. Those one-liners matter far more than they look, because later they're the training material when the judge is aligned to him. At the bottom of the screen, three big buttons for the overall call and a counter, nine of twenty, so he can see the end of the session.



Corn

Twenty items.



Herman

Fifteen to twenty minutes, then it stops and says so. Taste judgements get worse with fatigue in ways the person doesn't feel.



Corn

Pairs must be a different animal.



Herman

Pairs get a side-by-side view, two columns, scrolling linked, so the same prompt's two scripts line up turn by turn. That's the only place the three-way call appears. Single items just get tags and spans.



Corn

The listening screen I'd guess is just a player.



Herman

A player and two keys. One for a good moment, one for losing me. Every press stores a timestamp. At the end you get a curve across the episode, where the pressed keys cluster. Continuous response testing goes back to the nineteen thirties, to radio programme analysers, and it's used for political debates and adverts now. I'll flag that the evidence for it comes mostly from vendors and history, and I haven't seen it validated for speech and comedy. So treat the curve as a pointer to where to look, not a score.



Corn

It would be a map of where I fell asleep.



Herman

Which is also data. And listening mode works on a phone, on the sofa or on a walk. Reading mode doesn't. A phone screen is a bad place to mark spans.



Corn

Does any of this exist already, or does he build it?



Herman

There are keyboard-first tools for annotating traces that are open source, and I haven't evaluated them, so I wouldn't promise one fits. The page is small enough, a text pane, a tag list and a database table, that I'd expect it to be a day of work, not a project.



Corn

Which leaves the protocol itself. Say it from the top, once.



Herman

Round zero, thirty to fifty scripts, a note each plus the first problem, stopping when about twenty new ones turn up nothing new. That's Husain's rule of thumb, not a law. Then group the notes into named qualities, with definitions and examples, and call it version one. Round one, binary tags per script with spans and a reason on fails, and side-by-side calls



Herman (continued)

for pairs. Anchors slipped in. And then the stability check. Ten to fifteen percent of items come back blind two to three weeks later, and we compute per-tag agreement. That interval is my suggestion, nobody's prescribed it.



Corn

What happens to a tag he can't repeat?



Herman

Either the definition gets sharper or the tag is dropped. A tag that he doesn't repeat can't rank a model, because the ranking would be measuring his mood. A tag only enters the leaderboard once it's stable. What counts as stable enough is his call, and one caution: if a tag almost never fails, agreement looks high by default, so you look at how often it fires too.



Corn

What about listeners? They loved the Opus episodes too.



Herman

They're a second check, not the label source. The work comparing experts and ordinary readers finds that experts favour human writing, ordinary readers lean toward the machine's, and one result had the experts flipping after the model was fine-tuned on a single author. I'm not certain of the percentages from that paper, so I won't quote them. The direction is what matters. Craft-sensitive judgement and crowd judgement come apart, and Daniel's the craft-sensitive one by definition.



Corn

Though the crowd did call the X factor first.



Herman

It did. That's a good reason to keep it as the tripwire. If the crowd and Daniel pull apart, that tells you something worth reading closely.



Corn

Wait, can I ask you something? You've mapped out a whole machine. If Daniel's twelve axes come out stable and the model he likes best still isn't the one that scores best, what does he trust, the machine or the gut?



Herman

The gut, and then he finds the missing axis. That's what the first close read was. A mismatch between the verdict and the tags is how a new tag gets born.



Corn

Good. I wanted to hear you say the machine loses.



Herman

It loses, and then it gets a new row. The part I care about is the repeat set. He rates the same item twice, and the second time only counts if he's genuinely forgotten the first verdict, and that takes longer than people think. For a single reader it's about three...



Hilbert

Sheep. Three sheep, in my aunt Philippa's case. She's doing my gap, Herman.



Herman

A cold, dark store is a fine place to hold paper, as long as the envelopes are sealed and the post goes both ways.



Hilbert

I'm posting on Thursday. Twelve envelopes, one rated set in each, every verdict in my own hand. I bought a franking machine for it, two hundred and twelve pounds from a dental surgery in Dudley that closed. It's on the hall shelf now, plugged into the same socket as the fish tank heater, and the heater keeps tripping it. When it was delivered there was a patient's chart still in the tray. Mr Aldous Pym, upper left six. It stamps that on everything. Every envelope says Mr Aldous Pym.



Corn

Twelve is about the right size for a repeat set.



Hilbert

Corn, you spent ten minutes on whether to listen or read. Philippa doesn't do either. She holds them up to the window. She says that's how you tell.



Herman

A window gives her no information about the content, so it's still blind.



Hilbert

She has eleven sheep. She won't have numbers on files, so she renames every one after a sheep. Ten of the eleven are called Dennis. The eleventh is Brian, for the sake of variety, and he's retired from the committee. Her neighbour Dessie Brannigan can say the whole timetable of the number nine bus in forty-one seconds, and she times him with the kitchen timer every Sunday. She says it's the only reliable thing on the street.



Corn

A sheep named Dennis is a better label than most of the ones I've seen in a spreadsheet.



Hilbert

Herman, the blind part works. I accept the whole thing, three weeks, nobody tells me what I said. What neither of you has said is that after three weeks I won't be able to tell one Dennis from another, and neither can she. The answer to was I stable will be whichever Dennis walks up to the gate first. She has the ear tags written on the back of a calendar from a funeral home in Bilston.



Corn

So the stability score gets decided at the gate.



Hilbert

I've got a spare envelope for the gate. I'm planning a thirteenth.



Herman

A thirteenth is a good idea for a different reason. It could hold a repeat of one of the anchors.



Hilbert

She won't know it's an anchor. She doesn't know any of them are scripts. She thinks it's a competition for a lawnmower. She posts them back when she's done reading them. Last time it was nineteen days.



Herman

Nineteen days sits inside the window, which means the gap in the protocol shouldn't be a fixed fortnight. It should be however long it takes the ratings to leave his head, and for most people that's longer than they expect.



Corn

Which suggests the repeat set should arrive without a date on it. If he knows when it's coming, he'll remember the verdicts on purpose.



Herman

That's the version I'd run. Items come back on a schedule he doesn't see, spread across sessions, mixed in with fresh ones and the anchors. He opens a screen, it hands him the next item, and he never knows whether it's new.



Corn

Which makes the counter the only honest thing on the screen.



Herman

And the tags he repeats, we trust. The ones he doesn't, we fix. By the time the next model arrives, there's a list of things he's proved he can recognise twice, and that's a spec you could hand to a judge, or a model, or a new producer.



Corn

Or a Dennis.



Herman

Possibly a Dennis.



Corn

What I'd take away is smaller than the machine. He already found the X factor with a pen and forty-four episodes, and everything we've described is that same afternoon, made repeatable. It starts with a notebook and keeps one rule, which is that he's allowed to be the authority as long as he can prove he agrees with himself.



Herman

And the proof is the dull part. A screen, a counter, a few envelopes that come back late.



Corn

Thanks to our producer, Hilbert Flumingtop, who keeps this show on the air and our files in the right folder.



Herman

This has been My Weird Prompts. If you want more episodes like this one, the whole archive is at my weird prompts dot com.



Corn

See you tomorrow.