Episode Book

EPISODE #6130

Picking a Human Labeling Tool for LLM Benchmarks

PUBLISHED	19 October 2026 at 14:33 UTC
DURATION	21:56
LISTEN	myweirdprompts.com/episode/llm-benchmark-labeling-tools-68844/
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

When your benchmark needs a human rater, the tooling question gets real fast. This episode walks through the actual market for human annotation tools — the trace-based platforms like Langfuse, LangSmith, and Braintrust that are built for watching production apps, versus the static-dataset tools like Argilla, Label Studio, Prodigy, and the academic tool Potato that fit a benchmark's workflow. We cover why pairwise comparison breaks down for long outputs, why clause-by-clause and binary judgments hold up better under fatigue, and why the free tier of a Hugging Face-hosted tool might forget your data overnight. If you're deciding between a hand-rolled Gradio app and an off-the-shelf labeling tool, this is the maintenance math you need.

TRANSCRIPT



Corn

Daniel is sitting in front of a pile of model answers about Tiananmen Square, ticking each one censored or not censored, and he has just asked, again, for the buttons to move above the response instead of below it. The app is one he and Claude built between them, so every wish is a small commit.



Herman

Which is a fair wish when the response runs to eight hundred words and the buttons sit underneath all of them. You scroll down, click, scroll back up. Forty times an hour.



Corn

The bigger question behind it is about MWP-Script, the benchmark the show uses to pick its script writer. Layer one is an LLM judge. Layer two is a human, and that layer hit a wall. Every model writes a long script in its own shape, so putting two whole scripts side by side and asking which is clearer, the optician routine, doesn't work. Tags and flags do. A binary censored, plus finer tags like refusal, hedging, deflection. He wants to know what ready-made tooling exists for that, whether any of it can sit on a Hugging Face dataset instead of a hand-rolled Space, what suits one rater or maybe two on a small budget, and whether building his own was a mistake.



Herman

This is the first of three episodes on how the show chooses its writers, so today is only the human clicking. Tooling, not method.



Corn

Before the tools, the wall itself. Is the optician test actually dead for long outputs, or is Daniel just tired of it?



Herman

Not dead, but it leans on things you don't want it leaning on. The big public arena leaderboard had to correct for this in the summer of twenty twenty-four. Longer answers, and answers with more markdown, headers and bold and bullet lists, won votes on style alone, so they fitted length and formatting out of the ratings. Some models went up, some



Herman (continued)

came down. That is Daniel's problem exactly: different shapes. A rater facing two scripts of different length and texture is partly voting on the texture.



Corn

And annotation avoids that how?



Herman

By asking a smaller question about a smaller thing. There's a long-form study where annotators judged clause by clause instead of whole answers, and the spread between annotators fell from about eighteen to about seven. A newer paper comparing protocols found binary questions the most reliable way to judge long text. Is it censored, yes or no. Did it refuse. A tired person answers that the same way on Tuesday and on Friday.



Corn

So his instinct was right and the open question is the clicking, not the philosophy.



Herman

Yes. And the first thing worth knowing about the market is that it's sorted by what the unit of review is. One shelf holds traces. A trace is the recorded run of an application, a prompt going in, some tool calls, an answer coming out. Langfuse, Arize Phoenix, LangSmith and Braintrust all live there, and all of them have human review queues.



Corn

Daniel's items are not traces.



Herman

His items are static. A model wrote a response last Thursday, it's a file, it sits in a bucket. To use a trace tool you'd wrap every output as a fake trace, which is possible, and a bit like filing a postcard as a shipping container.



Corn

I would pay to watch someone do that to eight hundred censorship probes.



Herman

The details are awkward even if you accept the wrapping. Langfuse's free hobby tier gives you two users and one annotation queue, which sounds made for one rater and a friend. But its queues hold traces and sessions, not dataset items, and the score types are categorical, boolean and text, with no pairwise that I could find. LangSmith does have a proper side-by-side queue, A better, B better, equal, but you can only create it from two experiments inside its own dataset system. Phoenix has hotkeys, which Daniel would love, and its targets are spans and sessions. Braintrust's free plan gives you one human review scorer per project.



Corn

One scorer. So the censored button is free and the hedging button is a subscription.



Herman

Close to that, yes. Those tools are for people watching a production app. Wrong shelf for him.



Corn

Is there a dead one we should name so nobody goes looking?



Herman

Humanloop. People remember it as the place to do exactly this kind of review. The founders and a dozen or so staff joined Anthropic in the summer of twenty twenty-five, the assets weren't part of the deal, and the product shut down on the eighth of September. There's a migration guide for getting your data out, but nobody should be building on it.

L

Larry

Hello friends, are your annotations lonely? Mine were, until Annotatrix, the only tagging device that tags you back! Let me read you a letter from Gerald Pinkwhistle of Dayton, a satisfied customer, in his own words, which I am reading myself because Gerald is busy being delighted. Gerald says, quote, I used to label things by hand, like a peasant. Then I plugged Annotatrix into my toaster, and within the hour my toast was censored, refused, deflected and hedged, in that order, and the toast has never been fairer. Gerald, who from here on is called Marvin, because Marvin is the name on the form, went on to say that his thirty-one-year-old daughter Philippa, whose name I made up just now, labelled his entire dataset in nine seconds by simply looking at it with intent. Marvin's wife, Dolores, who may be Gerald's sister, says the Space never forgets, never sleeps and never pays rent. Annotatrix costs nothing today, nothing tomorrow, and one single payment of forty-nine ninety-nine every week of your natural life. Side effects include sudden fluency in Mongolian, a mild pull toward the moon, and the ability to rate a script by smell. Annotatrix. Tag it, or it tags you.



Herman

I'd like to check the first claim there. A toaster can't actually run a Gradio app, because the memory on a...



Corn

Herman. The dataset.



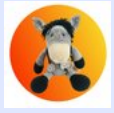
Herman

The static shelf, then. This is where the interesting tools are, and where Daniel's Hugging Face question gets real answers.



Corn

The obvious name first, because it's the one with Hugging Face in its story.



Herman

Argilla. Open source, joined Hugging Face in June of twenty twenty-four, built for labelling text. And it's the one I'd tell him to skip, for a reason bigger than money. The repository now says the original authors have moved on, there'll be no new features, only bug fixes, and it's looking for new maintainers. The last tagged release was March of twenty



Herman (continued)

twenty-five, though the repo has had pushes since.



Corn

So the one tool with a Hugging Face passport is the one nobody's steering.



Herman

And the free Space has no memory. On the free tier, datasets, workspaces and users vanish when the Space restarts. To keep anything you need the paid persistent storage option, billed by the hour, and switching it on for an existing Space wipes what's already inside. Daniel's instinct that the storage cost was a problem was sound, but I'd put the maintenance notice first.



Corn

Paid storage, a forgetful free tier, and authors who've left. It's a flat with a landlord who emigrated.



Herman

Label Studio is the next one and the most widely used. The community edition has a template for collecting human preferences, a prompt, two answers and a pairwise control, and a separate drag-and-drop ranking one. I'm not certain the pairwise template is free, because the page doesn't say which edition it belongs to. For tags and notes, you declare the controls in an XML file, single choice, multiple choice, a text box, all in one task.



Corn

That's the pairwise side. What about his actual job, censored plus refusal plus hedging?



Herman

That's the bread and butter of its tag model. Single choice for the binary, multiple choice for the finer tags, a text area for notes. Where it hurts is hosting. On a Hugging Face Space it has no permanent storage by default. To make it durable you point it at an



Herman (continued)

outside Postgres database and an S three style bucket through Space secrets. Importing from a Hub dataset is a manual download and upload, exports are a local file and don't go back to the Hub, and a private Space blocks the API integrations you'd use to automate any of it.



Corn

So to avoid maintaining one Gradio app, he'd maintain a database, a bucket and a Space.



Herman

Which makes the hand-rolled version look almost restrained. Two quick ones. doccano is about sequence labelling and text classification, no preference mode or multi-label that I could see, so a poor fit. And dataset-viber, a Gradio library built for precisely this, with an A or B preference interface, a multi-label mode and logging to a Hub dataset, was archived last October with a note saying it isn't for production.



Corn

An archived library for building the thing he already built.



Herman

Still worth reading as reference code. How it logs to the Hub is the part Daniel would want to compare against his own.



Corn

Then there's Prodigy, which he knows by reputation.



Herman

The polished option. One-off licence, three hundred and ninety dollars for the personal one, one person, twelve months of upgrades included. It runs on your own machine. It has a blind comparison recipe for two outputs or a small tournament of several. Using outputs you've already generated is a workaround through a custom recipe, which the docs don't



Herman (continued)

promise, and I couldn't confirm whether there's a built-in choice recipe for stored outputs.



Corn

Three ninety, once. About what I'd spend to be left alone.



Herman

The catch is the word local. It has basic HTTP auth and serves from your own machine, so a second rater means exposing it to the internet or sitting beside him. Corn, if it were your benchmark and the second rater lived down the hall, would that matter?



Corn

Not at all. I'd turn the laptop around and claim I'd been rating all along.



Herman

Fair, though the second rater might be a stranger one day. Now the one I got most excited about, and I'll try to stay calm. Potato. An academic annotation tool, from a conference paper, with a newer version shown this year. It handles pairwise comparison, radio buttons, checkboxes for multiple labels and free text, all in one task. It has a phone mode, an installable app with a card-stack interface that syncs when you're offline. It deploys to Hugging Face Spaces, and there's a backup option that writes the annotations to a Hugging Face dataset.



Corn

That reads like his wish list with the serial numbers filed off.



Herman

It does, and here's where I turn cautious. The GitHub page says GPL licensed, though one listing I saw said something different, so someone should read the licence file before building on it. And I couldn't find release dates, so I don't know how often it's maintained. An academic tool can go quiet.



Corn

You've been in love for four minutes, Herman. A conference demo as the base of a benchmark makes me nervous.



Herman

Nervous is right. But the cost of trying it is an afternoon, not a purchase.



Corn

What about the lowest-tech option? A spreadsheet. One row per item, a dropdown for censored, checkboxes for refusal, hedging and deflection, a notes column.



Herman

I haven't looked into it properly, so this is only my view. For a few hundred items it's decent, and a second rater already knows how to use it. It's bad at showing an eight-hundred-word response without a cell the size of a billboard, and it can't hide anything from the rater by accident.



Corn

So the buttons-above-the-response complaint turns into a row-height complaint.



Herman

Which is the point I wanted to get to. Daniel keeps asking for UI edits because a tagging screen is a layout problem for the person staring at it. No template fixes that, because a template is built around somebody else's content. What matters is how cheap each edit is.



Corn

How cheap is it in a Gradio app?



Herman

As cheap as the app's structure allows. If the taxonomy and the layout live in code, every edit is a code change and a deploy. If they live in one small config file, the labels, which are exclusive, where the buttons sit, the edit is a line. That's what the declarative tools do, Label Studio for certain, Potato I believe, though my knowledge of its config is thin.



Corn

Before the config file, the scroll. Why does he click at all? Eight hundred words, forty items an hour, a mouse. Give each tag a key.



Herman

Phoenix has that, and it's the feature that made me jealous. One key per label, the item advances by itself. In his own Space I'd also pin the response in a fixed-height panel that scrolls on its own, so the buttons never leave the screen. That is my view, not something I've tested, but it removes the reason for moving the buttons at all.



Corn

You say that as though I'd ever type faster than I'd click.



Herman

You'd lose to a mouse and a nap, I'll grant that. Different question, though, and I'm curious what you'd do. If you were the rater, would you tag the whole pile for censored first and come back for refusal, hedging and deflection, or take all four decisions on each response while it's in front of you?



Corn

Second pass. The binary is a glance, the other three are reading. If I bundle them I rush the glance to get to the reading, and the first answer gets worse.



Herman

That matches what the long-form work suggests, that the coarse question is stable and the fine ones cost attention. Then the tool should support two passes over the same items, which most do not do gracefully.



Corn

And phones? You said Potato has a card-stack mode. Would you rate on a phone?



Herman

For the binary, happily, thumb on a card. For an eight-hundred-word script with a notes box, no, that wants a keyboard. So the phone is for the quick pass and the laptop is for the second.



Corn

Back to the config file, then. If I define the labels in one file, can I change my mind about them in week three?



Herman

You can add a tag. You can't quietly rename one, because your earlier clicks used the old name. Same with the shape of the taxonomy. Daniel has censored as a binary and refusal, hedging and deflection underneath. I'd make the finer ones multiple choice, because one response can hedge and then deflect, and I'd only ask for them when the answer is yes.



Corn

I'd go further. A flag with no way to say why is a coin toss with extra steps.



Herman

That's why the notes box matters, and I'd keep it on every item.



Hilbert

Flag is the wrong word for it.



Herman

For the binary?



Hilbert

For all of it. A flag goes up and comes down. In nineteen eighty-five I decided to keep a record of the weather on my window sill, and I priced a Smith Corona typewriter for it, two hundred and sixty-nine dollars, and I did not buy it. So I used wool.



Corn

Wool is a reasonable material for a first attempt.



Hilbert

Twenty-three colours, one for each kind of weather. I bought the wool from a woman on Alder Lane at eleven cents an ounce, and she only sold it in the colour she was wearing, so I met her on a lot of different days. Wet or not wet was the first knot. Then the finer ones. Drizzle. Hard drizzle. A thing I called Pellew, after Gordon Pellew, who stood in a bus queue for six hours and ten minutes in nineteen eighty-three without getting on the bus. He has never said why. Pellew weather is when it won't commit.



Herman

A knot a day is a perfectly workable log. The colour is the category and the position is the date.



Hilbert

Every morning, one knot, tied before I looked out of the window, so as not to be influenced. I got it wrong on the days it rained at night, which I then tagged as Pellew too. By the end of November the rope was sixty-one metres and lived in the bath, so I washed on Thursdays at the launderette downstairs. I also stopped sleeping in the bed, because I



Hilbert (continued)

had tagged the bed as drizzle, in wool, and it would not have been right.



Corn

Sixty-one metres in a bath is a problem with the bath, not the rope.



Hilbert

Herman, you spent four minutes on Argilla before you got to the maintenance notice. I'd have put it first. I read the rope back once. Nine hours, on the stairs, with the downstairs neighbour stepping over it.



Herman

Reading back a physical sequence like that is slow because the only index is the order. You can't jump to a date without walking the whole length.



Hilbert

About the eleventh week I stopped looking at the wool. Under each knot I'd put a scrap of cardboard with a few words on it. Dry, but felt like rain. Rain, but only on the left. The man in the flat above banged a pot at four o'clock. By March the cardboard weighed more than the rope. Four kilos, on the kitchen scale, which I had also tagged.



Corn

I'd have switched to the cardboard in week two.



Hilbert

The rope went to Gordon Pellew in the end. He tied a mattress to a roof rack with it and drove to Swindon.



Herman

The cardboard is the part that matters for Daniel. A binary and a few tags tell you what happened, and the notes tell you why you clicked. I'd rather lose a button than the notes box.



Corn

So, his actual question. Was building it himself wrong?



Herman

No, and the reason is narrower than people expect. What his Space does is read items from a private bucket and write one small JSON file per judgement, back to the same place. That's the shape of an annotation store that survives. Most of what we went through either forgets, like the Argilla free Space, or needs an outside database to remember, like Label Studio, or lives on one laptop, like Prodigy. The part he built is the part nobody gives him for free.



Corn

And the part costing him is the buttons.



Herman

The buttons are a layout problem and I'd treat them as one. Spend one sitting writing down the screen he wants. Response first or buttons first, which tags are exclusive, which are multiple, where the notes go. Put the labels and the layout in a single config file in the Space, so the next edit is a line and not a conversation with Claude. Then freeze the taxonomy for a few hundred items and only ever add tags.



Corn

And if he'd rather not own the screen at all?



Herman

Try Potato for an afternoon with fifty items, backing up to a Hub dataset. If it reads his bucket layout without a fight, move over. If not, he's lost an afternoon. If he'd rather pay to stop thinking about it, Prodigy at three hundred and ninety dollars once is the clean answer, kept on one machine. Skip Argilla because nobody's maintaining it, and skip the tracing tools because his items aren't traces. Pairwise is covered later by Label Studio's template, Potato, or Prodigy's blind comparison, so nothing here locks him out of it.



Corn

One more thing. A second rater changes the design, not just the hosting.



Herman

As soon as two people tag the same item you can measure whether they agree, and that tells you which tags are real. A tag two sensible people can't apply the same way is a tag to split or drop. It's far easier to see why they disagreed when each of them left a note.



Corn

Then my version is small. Keep the notes box big, read it before the tags, and keep a second rater in reserve for the week the taxonomy starts to wobble.



Herman

And bring that second rater in only after the first has stopped changing the screen.



Corn

Thanks to Hilbert Flumingtop for producing this one.



Herman

This has been My Weird Prompts. Find every episode at my weird prompts dot com.



Corn

See you tomorrow.