Episode Book

EPISODE #6129

How to Build an AI Preference Dataset

PUBLISHED	19 October 2026 at 14:33 UTC
DURATION	22:35
LISTEN	myweirdprompts.com/episode/ai-preference-data-pipeline-08074/
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

Before a reward model can learn anything, someone has to build the dataset underneath it. This episode digs into the unglamorous machinery of pairwise preference collection: how candidate scripts get generated through batch inference, how pairs get chosen so the comparison graph stays connected, and why the aired script makes an awkward contestant. It also covers Bradley-Terry scoring, the cost of long-text comparisons, and a clever fix — aligned continuations — that cuts one script into shared prefixes so raters judge writing instead of plot.

TRANSCRIPT



Corn

Somebody has to build the A and the B before anyone can click either. That's the part Daniel wants to pull apart. He's collecting human preferences on the show's scripts, the simple kind, where a rater sees two responses and says this one or that one, like the optician flipping lenses. And his question is everything upstream of the click. How do you make the candidates, how do you decide who faces whom, how do you hide who wrote what? Plus, what's this simple form of feedback even called?



Herman

Name first, because it's the shortest answer. There isn't one official term. The papers say pairwise comparison or pairwise preference, the RLHF people say comparison data, and plenty of them say forced choice because the rater can't abstain. The optician version has a proper name in psychophysics, two-alternative forced choice. I haven't seen that term in the language model papers, so I'd keep it as the ancestor rather than the standard.



Corn

Forced choice sounds like something a hostage negotiator says.



Herman

It does. But pairwise preference is safe, and every paper will nod at it.



Corn

And this is part two of three, I gather. Last time was the tagging tools, refusal, hedging, deflection, labels on single outputs. Daniel says that's where he's headed long term.



Herman

Which is a fine instinct. Tagging a single output needs no partner. Pairs need a partner, and that's where all the engineering is hiding. Whatever the reward model does with the labels afterwards is a separate story. Today it's strictly the data going in.



Corn

His case first, then. The script writer runs through OpenRouter and Inspect AI. The builder takes every two usable scripts written for one prompt and pairs them, adds the aired script if one exists, and drops any model that had a technical failure.



Herman

Let me pull one design choice out before the machinery, because everything else follows from it. There are two ways to fill a pair. You sample twice from the same model and compare the two draws. That's on-policy, and it tells you which behaviours of that one model to reinforce. Or you take one draw each from different models. UltraFeedback did that, a pool of around seventeen models and four responses per prompt. Chatbot Arena does it live. That's the benchmarking design, and that's Daniel's.



Corn

This is my own synthesis, though, not a quote from a paper, right?



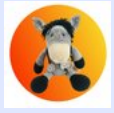
Herman

Yes, the framing is mine. The facts under it are the papers'.



Corn

So he's asking which writer is better, not how to train one.



Herman

Which means different models, same prompt, and the sample has to be fair to every model. That's why the failure rule bothers me. Dropping a broken output is obviously right. Dropping the whole model for that prompt, I'm less sure.



Corn

Because a failure isn't random.



Herman

Right. If a model times out on the long, strange prompts, and you remove it from exactly those, its record is built from the easy ones. It looks better than it is. I don't know how often that happens in his data. I'd check before trusting the leaderboard.



Corn

I'd check it, and then do nothing about it.



Herman

Corn, checking is the whole ask.



Corn

I'm a sloth. Checking is already a lot.



Herman

There's a cheaper way out anyway. Inspect can repeat samples with its epochs setting, and its eval set mode retries failures, ten times by default, and reuses what already finished. So the failure he drops today could be a retry tomorrow, with the model staying in.



Corn

Now how do you make a hundred scripts without staring at a progress bar?



Herman

Batch inference, and there's news. OpenRouter has a Batch API now. You send an array of requests, each with an id you choose, you get an accepted response straight away, you poll, and results come back keyed by your ids. It bills at roughly half the normal price with a twenty-four hour completion window.



Corn

Twenty-four hours is a long time to wait for a paragraph.



Herman

For this job nobody's waiting. The catches are small. The model has to be offered in its batch form or the request bounces. One provider per batch. No streaming. Results are kept about thirty days. I'd read the live documentation for the exact endpoint before building on it, because the details have been moving.



Corn

Do you trust that they all finish?



Herman

Providers set their own expiry, and some run longer than a day. So you build in the same thing as before. Collect what came back, find what didn't, resubmit only the gaps.



Corn

Sensible. Generation is the cheap part, then.



Herman

It is. Human minutes are the expensive part. You can afford to regenerate until every model has a clean output for every prompt.

Larry

Tired of choosing? Introducing the Prefer-O-Matic, the machine that prefers things so you never have to! Results in seconds, guaranteed, no waiting, no batches, no humans, just pure preference! Order in the next ten minutes and the price drops from nine hundred ninety-nine dollars to four ninety-nine! But wait, it's three forty-nine now, and we'll throw in a second Prefer-O-Matic, a lanyard, and a goat! Two goats! The price is now one dollar, and the lanyard is a different lanyard, and every model is paired with itself for maximum harmony, but hurry because this offer ends



Herman

Seconds is not how a batch works. The window is up to twenty-four hours, and the saving is about half. You trade speed for money, that's the whole deal. And pairing a model with itself is a real design, it just measures that model's own spread, not a ranking between models.



Corn

Let it go, Herman.



Herman

I'm letting it go. I'm only saying the goat is not a batch parameter.



Corn

Back to who faces whom. Four or five models. Just do every pair?



Herman

At his scale, yes. N models give n times n minus one over two pairs, so five models give ten. It also keeps the comparison graph connected, and ranking math cares about that. If model A never meets model D, directly or through a chain, you can't place them against each other.



Corn

And bigger?



Herman

InstructGPT showed labelers four to nine responses and had them rank the lot, which gives K choose two comparisons from one sitting. Rater effort grows roughly in a line, pairs grow with the square. It's the cheapest trick in the field.



Corn

So why isn't Daniel ranking five scripts in one go instead of ten separate A versus B jobs?



Herman

What would you do?



Corn

Nobody reads five scripts. My guess is the rater skims numbers three through five and calls it a day.



Herman

That matches what the long-text studies say, annotators skim long passages. Ranking works when the items are short. I'll come back to length because it's the big one. Arena is the other example. They don't sample pairs uniformly. They lean toward models of similar strength, since a mismatch tells you almost nothing. In their simulation, getting a given precision on the win matrix took roughly forty-four hundred adaptive comparisons against sixty-eight hundred random ones.



Corn

Sixty-eight hundred. Daniel will get there around the heat death of the show.



Herman

But for the single score per model the gap was small, about sixteen thousand against seventeen thousand. At his size, all pairs with some repetition is fine. Clever sampling is for people with a crowd.



Corn

And the score itself is Bradley-Terry.



Herman

Each model gets a latent strength, and the chance that A beats B depends on the gap. You fit it by maximum likelihood. A tie counts as half a win each. Arena switched from Elo to Bradley-Terry partly because the order the votes arrived in stopped mattering.



Corn

One thing about the aired script. It's in every pair when it exists. Isn't that a bias of its own?



Herman

It's a different kind of entry. The aired script went through polish and a text to speech pass, and it was picked from whatever the pipeline produced that day. So when a raw model draft meets it, you're measuring the model against the show's shipping standard. I like it as an anchor, but I'd keep it in its own column and not fold it into the model table.



Corn

A benchmark that's also a contestant.



Herman

And a contestant that chose its own prompts, in a sense, since the prompts are the ones that aired. That could lean easy. Not a flaw, just something to remember when the numbers come in.



Corn

What about one sample per model per prompt? That feels thin.



Herman

It is thin. One draw from a model is one roll, and a model that writes a great script one time in three looks like a coin flip on a single sample. You want a few draws per model per prompt. Then you can pick the pairs from the pool, or let the same pair type repeat across different draws. Inspect's epochs make that a setting, not a project.



Corn

Does anybody say how many draws?



Herman

Not that I've found. UltraFeedback used four per prompt, InstructGPT used four to nine. For a house eval I'd start at two or three per model and see how much a model disagrees with itself.



Corn

How much it disagrees with itself is a number I'd like.



Herman

It is a great diagnostic. If a model's two draws beat each other half the time, the rater has nothing to learn from that pair, and you've also measured the noise floor for the whole table.



Corn

Next worry. Ten pairs from one prompt. Does the rater see them together?



Herman

InstructGPT treated all comparisons from one prompt as a single unit when training, because they're correlated, and I'd think the same about collection. The prompt is the unit, not the pair. Ten pairs from one prompt are not ten independent votes. So when you count how much data you have, count prompts.



Corn

And scheduling them? If the rater sees all ten in a row they'll be bored by the fifth.



Herman

Spread them. Interleave prompts so no one sees the same prompt twice in a row. Otherwise the rater starts to remember what a previous model did with the same setup and carries it across.



Corn

Memory, or fatigue.



Herman

Both. And a rater who has just read the winner of a pair will be hard on whatever comes next.



Corn

Okay. Whole scripts. Where it gets painful.



Herman

Two models diverge inside a minute. Different examples, different jokes, a different route through the topic. The rater isn't comparing two answers anymore, they're comparing two different shows, and the vote ends up resting on whatever they were tired of that morning.



Corn

And it's a long read for one click.



Herman

Annotators skim, and agreement on long comparisons comes out low. That's what Daniel's aligned continuations are aimed at. Take one shared script, cut it at roughly fifteen, fifty and eighty percent, hand every model the same text up to the cut, and have each write the next three hundred words. Then compare those blind.



Corn

Everyone starts from the same place and the rater only judges the next step.



Herman

A much better question. Three hundred words is about two minutes aloud, the rater can read all of it, and both texts do the same job at the same point. Differences are about writing, not plot.



Corn

Where does it break? You're about to tell me.



Herman

Four spots. First, whose script is the prefix? If it's the aired script, one model's voice is baked into the setup, and models that imitate well may beat models that write well. Second, the cut. Mid-sentence means the task is repair, so cut on a turn boundary. Third, length. If one model writes four hundred and another two hundred and fifty, the longer one tends to win.



Corn

So cap the length.



Herman

Instruct it, then discard the overshoots. I wouldn't trim, trimming mangles the ending. Fourth, the three cut points are three different skills. At fifteen percent a continuation is setting up, at eighty it's heading for the close.



Corn

I'd have said pool them. More data.



Herman

Pool them and a model that's strong at endings and weak at openings looks average. That's the one thing he'd want to know.



Corn

Hm. Fair. Keep them apart, pool for an overall.



Herman

And it fits his existing item kinds. The opening, around four hundred words, is already a continuation from nothing.



Corn

I have a worry about the prefix. If the shared script came from the same model that's one of the contestants...



Herman

Then that contestant is continuing itself. Yes. I'd take the prefix from the aired script and accept that whoever wrote it has a small home advantage. Or pick the prefix from a model that isn't in the contest.



Corn

Or just look at whether it matters. If the home model wins more at the cut it wrote than elsewhere, you'd see it.



Herman

That's better than my version, actually. Test the bias instead of arguing about it.



Corn

Here's what I'd actually ask the rater. Which would you rather hear on the show?



Herman

What goes under the two texts matters. What do you put under the two texts? If you say which is better, you get taste, mood and length, all tangled. If you split it into voice, clarity and accuracy, you get more signal but more work, and you're half-way to the tagging approach again. A single question keeps the pair fast. What would you pick?



Corn

One question, but the right one. Which would you rather hear, because the thing is going to be read out loud, by me, with my timing.



Herman

Your timing is a variable I'd like to remove from the dataset.



Corn

I'd like it removed from the show. It keeps writing me into a nap.



Herman

There's one more reason to keep the question single. Agreement among raters on this sort of judgement isn't high. InstructGPT's labelers agreed with each other around seventy-three percent of the time, and a recent story benchmark found its best off-the-shelf model judge at seventy-three percent agreement with humans, trained reward models at seventy-eight. Those are different measurements, but the lesson is the same, nobody gets to ninety.



Corn

So a rater who disagrees with you is not necessarily wrong.



Herman

And a single rater's vote is a small dose of one person's taste. For Daniel's case, that's fine, because the taste is the point. It's the show's taste. Serving is where raters get misled without noticing, though.



Corn

Go on, then.



Herman

Side assignment is random per pair, and the mapping from side to model sits somewhere the rater's screen never touches. Strip anything that gives the source away. Formatting habits, a signature first line, a model that always italicises the host names. Arena drops conversations where a model names itself, same idea.



Corn

What about the ten percent he repeats with the sides swapped?



Herman

That measures position bias. People lean left, or right, or toward whatever they read last. Show the same pair again flipped and see if the rater flips with it. The judge literature has a convention that a win only counts if it survives both orders, otherwise it's a tie. Ten percent gives a read on each rater's lean without doubling the work.



Corn

Would you let them say tie?



Herman

Tie, and also both are bad. Arena has both buttons. A forced pick between two broken texts produces a confident label that means nothing.



Corn

And at the end he has prompt, chosen, rejected.



Herman

Every decisive vote becomes one triple, winner as chosen. Ties go into the ratings, half a win each, but never into triples, since there's no chosen side. Model names ride along in the metadata and never reach the rater.



Corn

What does one rater cost, in minutes?



Herman

I can't give you a number I trust. It depends on how much of three hundred words a person really reads.



Hilbert

How many words does a person read of the left-hand one before deciding the right-hand one is worse? I put forty kettles through the man in flat nine, and it was nowhere near three hundred.



Herman

Kettles?



Hilbert

Two bits of paper, one folder. Mr. Teodor Vask. I slid them under his door during the part about batches, Herman, and he answers through the letterbox. He holds each one up to the stairwell light.



Corn

He read forty of them in that time?



Hilbert

Not him. The tortoise does the picking. Inspector Halloran. He sits on the one he prefers, and Mr. Vask writes down which. The tortoise sat on the right-hand one thirty-eight times.



Herman

Then print the two sides on the same stock and put the tray on something that turns.



Hilbert

Mr. Vask had a lazy Susan. Same result, so then he said it was the paper. I'd printed the right-hand ones on Dusty Apricot, because the grey ran out and I had a ream of Apricot left over from the council by-election.



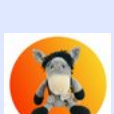
Corn

Apricot paper is a tell no formatting rule will catch.



Hilbert

The tortoise is paid, Corn. Eleven tins of sardines a session, in advance. Mr. Vask keeps three for invigilating. He wears a lanyard that says Chief Examiner, laminated by the dentist on the second floor.



Herman

You'd want the tins in the budget as a flat fee, so the pay doesn't depend on which side he picks.



Hilbert

It doesn't. The tins go in the middle of the tray, Herman, same as your ten percent repeats, which you said like they were free. They cost two tins each. After the Apricot business I put everything on grey and the tortoise started picking the left, thirty-one times out of forty, and Mr. Vask filed a complaint with the building committee. The committee is Mr. Vask and a man called Lennart who lives in the lift.



Corn

A committee of two, one of them vertical.



Hilbert

Lennart upheld it. The lift stops between the fourth and fifth floor on Thursdays, so the ruling was read to the tortoise through the gap. Then Mr. Vask moved the whole operation to the roof, because the stairwell light was a forty-watt bulb and the tortoise had begun to prefer whichever page was brighter. The roof has a pigeon that has taken sides.



Herman

Light on one side is a leak in the same family as the paper. A flat grey page and an even light, and the leak is closed.



Hilbert

Eleven minutes a kettle. That's the egg timer, from the sardine to the sitting. Forty kettles is seven hours and twenty minutes, and the roof closes at six.



Herman

Eleven minutes per pair is a number I'd plan a rater budget around, whoever the rater is. Forty pairs a sitting is out of reach for most people anyway.



Corn

It moves the whole picture. A thousand pairs is a hundred and eighty-odd hours of someone's life. Nothing about generating the candidates costs anything like that.



Herman

Which argues for spending the effort on what goes into the kettle. Same prefix, same length band, same stock, blind sides. A rater's eleven minutes shouldn't be spent on a pair that was lopsided before they looked at it.



Corn

You said kettle.



Herman

I did. It stuck.



Corn

I'm going to say pilot run. Twenty or thirty kettles, one rater, a stopwatch, before any big batch gets paid for. You find out what a pair really costs and whether the right-hand side wins more than it should.



Herman

That's the sensible order, because everything else in the plan has a free knob except the human. Generation is half price. A model can be regenerated overnight. The rater is the one thing you can't re-run.



Corn

Misconception time. People assume more comparisons always gives a better ranking.



Herman

It depends on which ones. Thousands of pairs between two models with a huge gap teach you almost nothing. A couple of hundred between close neighbours teach you a lot. Margin and how well the graph is connected drive how fast a Bradley-Terry ranking settles, and I haven't found anyone with a magic number.



Corn

So for Daniel, the thing to watch is whether the pairs are worth the eleven minutes, not how many there are. That's my one forward thought. Make the next kettle the one you least know the answer to.



Herman

Thanks to Hilbert Flumingtop for producing. This has been My Weird Prompts. Part three, putting the model judge next to the human one, is coming up. You'll find every episode at my weird prompts dot com.



Corn

We'll be back soon.