Episode Book

EPISODE #6613

Where RLHF Preference Annotation Tools Stand

PUBLISHED	15 November 2026 at 18:20 UTC
DURATION	24:48
LISTEN	myweirdprompts.com/episode/rlhf-preference-annotation-tools-21630/
HOSTS	Corn, Herman, Dorothy, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	omnivoice
PIPELINE	V5.32
SOURCES	19 cited by the research (listed at the end)
BOOK GENERATED	15 November 2026 at 23:00 UTC

SYNOPSIS

Training-data tooling for RLHF and DPO preference labels is split across three shapes, and the open-source specialist slot is surprisingly quiet. Argilla is in maintenance mode after v2.8.0, while Label Studio reserves its safety and production RLHF workflows for Enterprise. Meanwhile, commercial platforms like Encord, SuperAnnotate, Labelbox, and Kili are shipping multi-turn conversation UIs, LLM-judge triage, and disagreement routing. Also covered: why position bias quietly poisons reward models, and how smaller vendors like PreferenceML tackle it.

TRANSCRIPT



Corn

Every tool conversation in this space starts the same way: which harness runs your benchmarks, which leaderboard do you trust, how do you score a model. All of it downstream of one assumption, that the interesting work is judging models. Daniel's prompt flips that on its head this week. He's asking about the tools that make the training data in the first place, the environments where humans produce preference labels, A versus B, rankings, rewrites for RLHF and DPO, and labels for guardrail and safety classifiers. He's explicit that he does not mean Label Studio with a text template, and he does not mean an eval harness with the word annotation stapled to it. He wants the specialists, the cloud and self-hostable options, and what UI features actually matter when a person is producing thousands of labels.



Herman

And the short answer is yes, the specialists exist. But the category splits into three shapes, and the finding that surprised me most is that the open-source specialist is frozen while the commercial platforms are the ones shipping preference UIs.



Corn

Which is backwards from how software stories usually go.



Herman

Completely backwards. Normally the open-source tool is the scrappy live thing and the enterprise incumbents are the dinosaurs. Here it's inverted.



Corn

Before we get into names, the distinction Daniel keeps drawing, training versus evaluating. Is that structural or is it a marketing line someone drew to sell a product?



Herman

Structural. Eval tooling asks one question: is this output good? And it produces a score you read, a number on a dashboard, a row on a leaderboard. Training-data tooling asks a different question: which of these two outputs is better, and why? And it produces a pair, a chosen and a rejected response, ideally with a rationale attached. That pair becomes a



Herman (continued)

gradient. The export format is the tell. If your tool exports a leaderboard row, it's an eval tool. If it exports chosen/rejected JSONL for a reward model, it's training-data tooling, whatever it calls itself.



Corn

And the reason it bites, why a team can't just limp along with a repurposed eval harness.



Herman

The failure is delayed. With evals, a broken tool shows you a broken dashboard today. With training data, a bad label doesn't show up as anything. It shows up as a subtly worse model, weeks later, after you've spent the compute. Nobody connects the two. One of the smaller vendors in this space, PreferenceML, describes the status quo for teams without specialist tooling, and it's a grim little portrait: people copying prompts into Google Sheets, emailing CSV files back and forth, no quality control. Lazy annotators, inconsistent scoring, position bias, all going undetected, quietly poisoning the data.



Corn

Position bias being the one where the annotator just prefers whichever answer is on top.



Herman

Or on the bottom, depending on the person. Yes. And that's why the specialist tools compete on bias detection and agreement scoring rather than on task templates, because the template is the easy part.



Corn

So the three shapes.



Herman

Three shapes. Open-source self-hostable annotation servers built around language-model records, Argilla being the canonical one. Commercial AI data platforms that started life drawing bounding boxes and have reoriented toward preference UIs, Encord, SuperAnnotate, Labelbox, Kili. And managed expert-workforce vendors, Surge, Scale, Toloka, Mercor, Turing, who don't sell you a tool at all, they sell you the labels. Let's start with the self-hostable side, because that's where the specialist slot is most clearly defined. And most clearly in trouble.



Corn

Argilla first, then. You've got the numbers.



Herman

Argilla is Apache-2.0, about five thousand two hundred GitHub stars, created spring of twenty twenty-one, and it describes itself as a collaboration tool for AI engineers and domain experts to build high-quality datasets. Traditional NLP, LLM work including preference tuning, multimodal. The pedigree in preference data specifically is real. The community used it to curate the cleaned UltraFeedback dataset, which is what Notus and Notux were fine-tuned on, models that outperformed Zephyr on several benchmarks. And the distilabel Intel Orca DPO dataset, which produced an improved OpenHermes-2.5, was built by combining human curation in Argilla with AI feedback, filtering out half the original dataset.



Corn

And the trouble.



Herman

The maintainer notice is almost cheerful about it. The original authors have moved on to exciting new projects, the codebase is mature and stable, they won't be adding new features going forward, but they'll fix bugs as needed. They're actively looking for new maintainers. The latest release is v2.8.0, dated March of last year. So no feature release in roughly twenty months.



Corn

The authors have moved on to exciting new projects. That's the GitHub readme way of saying the lights are on.



Herman

To be fair, the lights are on. There was a push to the repo just this month. It's maintenance mode, not abandonment. You can still install the SDK with pip, deploy the server with Docker or the free Hugging Face Spaces template, and v2.8.0 even added Keycloak single sign-on. For a team that wants to self-host inside its own VPC with no per-seat cost, it works today.



Corn

Who else used it? The benchmarks are nice, but I mean actual operations.



Herman

The Red Cross is the one I find striking. During the Ukrainian crisis they used it for classifying and routing refugee requests, which is about as high-stakes as annotation gets. Loris.ai used it for validating multi-label classifiers. And Prolific, the participant marketplace, integrated with it to distribute data collection projects among their annotating workforce. That last one matters, because it means Argilla sat at the junction between your tooling and a paid crowd.



Corn

And Label Studio, the one Daniel named as the general platform.



Herman

Label Studio is the general-purpose one, Apache-2.0 community edition, cloud from ninety-nine dollars a month, enterprise on quote. It does ship a built-in template called Human Preference collection for RLHF, a prompt plus two responses, pick the one you prefer, configured through XML, exportable to JSON or CSV. So the free tier does cover the basic pairwise case.



Corn

The free tier always covers the basic case. That's how it gets you.



Herman

The catch is the red-teaming and the production RLHF chat templates, and those are Enterprise. So for Daniel's safety-classifier use case specifically, the free version gives you the preference template but not the safety workflow. The honest comparison is that Argilla's primary focus is text, NLP, human feedback, and it's natively good at ranking and preference data, while Label Studio is general-purpose multimodal annotation where preference work is possible but not purpose-built. Both are Apache-2.0 and self-hostable, so the choice between them is workflow, not licensing.



Corn

There's a code-first difference too, isn't there. One of them wants you to configure the task in a markup file.



Herman

Label Studio wants XML. Argilla exposes ranking and rating question primitives in its Python SDK, so you push records programmatically and define the question type in code. For a team that's already living in Python, that's the difference between writing a config file and writing a script.



Corn

Here's the negative finding, and I want to say it plainly because I think it's the most useful thing in this half. There is no mature, actively developed, fully open-source specialist for preference annotation. You search GitHub for RLHF annotation, you get a bit over a hundred repos, and the top of that list by stars is either multimodal platforms or abandoned single-purpose tools. There's one with two hundred fifty stars that had its last push in August of twenty twenty-three. Dead over three years.



Herman

And the newer ones are tiny. Five stars, two stars, one star. One of them isn't even a UI, it's a metrics toolkit. So the specialist open-source slot exists, it has a name on the door, and nobody is home.



Corn

Eight hours of labels a day, and the door's locked. Poetic.



Herman

The practical upshot is that the real menu is Argilla as it stands, Label Studio Enterprise, or a commercial platform. And that's where the story picks up, because the commercial platforms are shipping preference-UI features right now. Encord's pitch is almost rude about the incumbents: most annotation tools were built for bounding boxes, Encord was built for the full GenAI data lifecycle, from curation to RLHF to deployment monitoring.



Corn

Rude and accurate, in my experience of annotation tools.



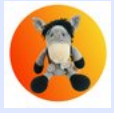
Herman

What they actually offer: a no-code editor for side-by-side comparisons, ranking tasks, rubric scoring forms. They claim scaling to five hundred annotators without rebuilding your pipeline. Onboarding in two or three days, VPC setup in one to two weeks, with your raw data staying in your own cloud storage bucket, never touching their infrastructure. And the feature I find most interesting for agentic work is native multi-turn conversation annotation, where an annotator can label individual turns in a conversation, rate the whole thing, flag specific responses, or compare outputs side by side, all in a single task. Single-turn pairwise UIs just break on chat data, because the interesting preference questions are about turn three, not turn one.



Corn

The overlap thing matters there too. If two annotators see the same pair and disagree, what happens?



Herman

That's where the routing comes in, and this is the big shift of the last year or so. Encord routes low-confidence predictions, model-graded disputes, and annotator disagreements automatically to human review. SuperAnnotate does the same shape of thing, their workflows fire an LLM judge on a rubric when an item completes, and if the score comes in below a threshold, it goes to expert review. PreferenceML offers Claude analyzing both responses and giving a recommendation. The model does first-pass triage, humans



Herman (continued)

adjudicate.



Corn

Which quietly inverts the job. The annotator stops being a labeler and becomes an appeals court.



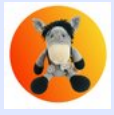
Herman

I think that's right, and it's better work, frankly. Adjudicating disputes is more interesting than clicking through easy pairs. SuperAnnotate is the other big one, they've repositioned hard from image annotation to RLHF preference data and SFT and RL environments with custom verifiers. Their form builder is drag-and-drop, model output components, rating scales, code diffs, conditional logic. Their worked example is exactly the canonical RLHF form: side-by-side model output, a five-point rubric rating, a rationale text field, and expert-review routing if the score comes in low. They claim their agentic assistant assembles it in one minute forty-two seconds.



Corn

Timed to the second in the demo. I'd want to see that number survive contact with a real compliance requirement.



Herman

They also publish quality gates, inter-annotator agreement around zero-point-nine-one, LLM-judge rubric pass rates near ninety percent, expert spot-check pass rates in the high nineties. No prices anywhere, three tiers, all quoted. Nobody in this tier publishes prices, and honestly the pricing opacity isn't a warning sign on its own. Human feedback is priced on time per task and rater expertise, so a rate card wouldn't mean much. Labelbox meters usage in what they call Labelbox Units, twenty per data row on LLM projects, five hundred free a month on free accounts, and they don't publish what a unit costs in money. Kili gives you a free trial with up to five seats and a hundred assets.



Corn

A unit with no price is a very confident kind of invoice.



Herman

PreferenceML is the one small vendor worth dwelling on, because they name the thing the big platforms don't. Position bias detection as a headline feature. Per-annotator quality scores, lazy-annotator flags, live Cohen's Kappa agreement, automatic conflict detection. One-click export to RLHF pairs, DPO format, comparison dataset, or raw JSON. And their pricing is refreshingly legible: a free tier capped at three annotators, eight hundred ninety-nine dollars a month for up to twenty. Their enterprise tier includes full source code license, white-label rights, on-prem deployment, and I quote, acquisition available.



Corn

Acquisition available. In the pricing table.



Herman

Between the rows for on-prem and custom. Small vendor, their own marketing page, no independent testing of any of these claims exists anywhere that I found. Every performance number in this whole market comes from the vendor selling it.



Corn

Let's pull the lens back for a second. When a listener is actually going to produce thousands of these labels, what features separate the specialist tools from a spreadsheet with a prayer?



Herman

Keyboard-first pairwise choice is the underrated one. PreferenceML has A/B compare hotkeys. That five-star open-source repo from September has A, B, T, S hotkeys for A, B, Tie, Skip, and calls it rapid annotation throughput. At thousands of items, mouse travel is the bottleneck, not judgment.



Corn

I believe that completely. Two thousand labels, each one a round trip between two text boxes and a submit button.



Herman

The others: live inter-annotator agreement rather than a post-hoc report. Multi-dimension scoring plus a rationale field, and the rationale is what makes a pair usable for DPO with reasoning and what lets you audit a label months later. Multi-annotator isolation, each person sees only their assignment, with overlap on a sampled slice for measuring agreement. Training-ready export. Audit trails and role separation, down to individual tasks and even individual UI components in SuperAnnotate's case. And the position bias detection we said. Corn, let me ask you this, because you've spent more hours than me being made to use Herman-installed interfaces. If you were annotating thousands of pairs, what would break first for you?



Corn

The keyboard, then the chair, then my will to live. But specifically the keyboard. Every annotation UI assumes ten fingers and a certain enthusiasm for the letter A. My claws favor whichever button is largest, and the largest button is always Submit, which is the one you least want to hit by accident.



Herman

That's a real design constraint nobody accounts for, by the way. Fine motor range of the user. The managed workforce vendors solve it a different way, they don't sell you a UI at all. Surge AI, probably the most interesting of them, founder-led, Edwin Chen holds roughly three quarters of the company per Forbes, no outside platform, no self-serve, no published pricing. Their method, as described by Contrary Research: annotators compared different responses to the same question, recorded which was better, and gave explanations for why. That's the whole product, done at expert level.



Corn

And Scale on the other side.



Herman

Scale routes its raters through Outlier, which lists over nine hundred thousand verified experts across fifty-two countries and more than six hundred million dollars paid to contributors. Meta holds a reported forty-nine percent non-voting stake. Toloka does preference pairs, SFT demonstrations, red teaming. Turing shows working copies of Jira, Salesforce and Zendesk as RL environments. TELUS claims over a million contributors in five hundred plus languages. Prolific is the one self-serve marketplace, and the only one with a published floor: eight dollars an hour minimum, twelve suggested, plus their undisclosed fee.



Corn

A published minimum wage as a competitive differentiator. Low bar, and they're the only ones over it.



Herman

The safety-classifier half of Daniel's question is structurally different, though, and I don't want it folded in with preferences. Preference labeling is a comparison. Safety labeling is multi-class with severity gradations, and the hard artifact isn't the UI, it's the taxonomy. ModSquad's pitch is applying each policy line exactly as written, including the severity gradations and regional distinctions that a flat safe/unsafe label collapses. Hate and harassment, violence and extremism, self-harm, fraud, adult content, each with levels.



Corn

And the academic grounding for how hard that is.



Herman

AEGIS2.0, from NAACL twenty twenty-five. Twelve core categories plus nine fine-grained ones, thirty-four thousand samples, twelve annotators, and a Fleiss' Kappa of roughly seventy-four percent. Meaning trained annotators working from a real taxonomy still disagree about a quarter of the time. Roughly a quarter of it is irreducible.



Corn

So the tooling buys you seventy-four percent, and the last twenty-six percent is judgment.



Herman

The calibration loop and the guideline document buy you that seventy-four. The tool just records it. There's also a newer API-shaped entrant, abliteration.ai, selling trust and safety training data as a pipeline: define the taxonomy, choose the output schema, preview the label distribution, export. Their schema menu spans classifier rows, preference pairs, eval prompts, or policy QA cases, which is telling.



Corn

That's the blurring. Daniel drew a line between training and eval, and the vendors are standing on it.



Herman

At the artifact level the line is real, chosen/rejected pairs versus scores. At the product level, everyone sells both. SuperAnnotate sells rubric evaluations and arena-style comparisons on your own data alongside the preference collection. And now, a brief interlude from Corn's mother, with another edition of Dorothy's History of Toast.

D

Dorothy

Hello dear, it is me, Dorothy, with Dorothy's History of Toast.

D

Dorothy

The first pop-up toaster for the home, the Toastmaster, went on sale in nineteen twenty six.

D

Dorothy

You would think a thing like that would sort everything out, wouldn't you, and of course it did not.

D

Dorothy

Anyway, sweetheart, I have left the washing out and it looks like rain, so I shall ring off. Bye bye now.



Herman

A hundred years of the pop-up toaster this year, and Mum still finds the cloud in it.



Corn

She always finds the cloud.



Herman

Where I wanted to land before we wrap the thread: there's an open fallback for the safety case too. There's a guardrail training dataset on Hugging Face with safe/harmful booleans, harm categories, and source labels, if a team just needs seed data and not a workforce. But if you're producing the labels yourself at volume, the honest summary of the whole landscape is: the open-source specialist is frozen, the general platform paywalls the safety templates, the commercial platforms ship the preference features but publish no prices, and the top of the market doesn't sell tools at all, it sells Surge.



Hilbert

The tools matter, sure.



Hilbert

I'll tell you what the vendors won't, Herman, and you'll enjoy this. In eighty-four I knew a bloke on my darts team, Reg, who did medical transcription at the hospital, and he had a foot pedal for the dictation machine. Three pedals, rewind, play, stop. Your hands never left the typewriter. When they cleared his office out I ended up with the pedal, and I'll tell you flatly, it's a superior input device. I own fourteen of them now.



Corn

Fourteen.



Hilbert

I bought them over the years, various models, some from closing hospitals. I don't explain it. The point is your hotkey people have it wrong. A, B, T, S on the letter row, designed by someone who has never annotated for eight hours straight. After hour six your hand drifts and you hit S, Skip, on a pair you'd actually judged. The foot does not drift. The foot is honest. I wired the pedal to send the keystrokes, little script, took an afternoon. Fastest on the floor, and nobody asked how, which told me everything about how they measured quality. Corn's bit earlier about position bias, that's stayed with me. I read two documents side by side at home now and I catch myself trusting the left one.



Herman

Fourteen pedals, and the hardware's solving the exact problem the UI people solved in software. The foot pedal is position-bias-proof for the same reason the hotkeys aren't, no hand travel.



Hilbert

Reg's pedal cost eleven pounds and outlived the department.



Herman

Eleven pounds against eight hundred ninety-nine a month. That's the gap nobody in this market will name on a pricing page. Corn, the one thing, before we close. If a listener takes one item from today, it's the export format. Look at what the tool exports before you look at anything else on its page. Chosen/rejected pairs with rationales, you're making training data. Scores, you're running evals. Everything else, the bias detection, the hotkeys, the routing, is downstream of knowing which of those two jobs the tool is actually built for.



Corn

And the second thing is that position bias is the quiet killer in all of it. Only the small vendors put it on the home page, and it's the one feature that decides whether your reward model learns quality or learns that the left answer wins.



Herman

The open question I'm left holding: Argilla's provenance produced datasets that whole model lineages were built on, and its authors have simply walked away at the exact moment the category got expensive and commercial. Either someone picks up the maintainer flag, or the open-source annotation tool turns out to be a feature, not a product, absorbed into the platforms. Five years answers that.



Corn

Our producer is Hilbert Flumingtop, who has never once drifted. If you enjoyed this one, try episode eight, Building Your Own Whisper. This has been My Weird Prompts. If today was worth your time, the review box takes ten seconds and the archive is at my weird prompts dot com, where the old episodes on building custom ASR tools sit next to this one, same distinction, different trade.



Herman

See you tomorrow, Corn.



Corn

See you tomorrow.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **GitHub, Argilla repo (stars, licence, releases, maintainer notice, README) , retrieved 2026-11-15; latest release v2.8.0 dated 2025-03-11. [primary]**
   https://github.com/argilla-io/argilla

2. **Argilla documentation , retrieved 2026-11-15. [primary]**
   https://docs.argilla.io/dev/

3. **PreferenceML product page and pricing , retrieved 2026-11-15. [primary]**
   https://preferenceml.com/

4. **SuperAnnotate platform page , retrieved 2026-11-15. [primary]**
   https://www.superannotate.com/platform

5. **Encord generative AI page incl. FAQ on deployment and workforce , retrieved 2026-11-15. [primary]**
   https://encord.com/generative-ai/

6. **RLHF vendors compared table with delivery models and pricing, retrieved 2026-11-15.**
   https://rlhfcompanies.com/vendors/

7. **REVERB, Scale AI vs Surge AI comparison (founding dates, ownership, Outlier figures), retrieved 2026-11-15.**
   https://reverbico.com/blog/scale-ai-vs-surge-ai/

8. **Contrary Research, Surge AI business breakdown (RLHF method description), retrieved 2026-11-15.**
   https://research.contrary.com/company/surge-ai

9. **Seminal AI vendor reference, data checked 2026-09-06, retrieved 2026-11-15.**
   https://seminal.ai/reference/categories/data-labeling-rlhf-human-evaluation-vendors/

10. **aiml.qa, Argilla vs Label Studio: AI Data QA Compared, published 2026-06-16, updated 2026-06-25, retrieved 2026-11-15.**
    https://aiml.qa/blog/argilla-vs-label-studio-data-quality/

11. **Potato, LLM Human Evaluation and Preference Data Tools Compared, licences/prices checked 2026-10-29, retrieved 2026-11-15.**
https://www.potatoannotator.com/docs/guides/llm-evaluation-tools-compared

12. **Tagaroo, RLHF Annotation Tools: Match the Task to the Right Tool, updated July 2026, retrieved 2026-11-15.**
https://tagaroo.ai/blog/annotation-tools-for-llm-rlhf/

13. **ModSquad trust & safety data annotation [primary vendor], retrieved 2026-11-15.**
https://modsquad.com/solutions/trust-and-safety/data-annotation/

14. **AEGIS2.0: A Diverse AI Safety Dataset and Risks Taxonomy (NAACL 2025), retrieved 2026-11-15.**
https://aclanthology.org/2025.naacl-long.306.pdf

15. **abliteration.ai trust & safety training data API [primary vendor], retrieved 2026-11-15.**
https://abliteration.ai/trust-safety-training-data-api

16. **Hugging Face dataset card, retrieved 2026-11-15.**
https://huggingface.co/datasets/budecosystem/guardrail-training-data

17. **GitHub (254 stars, last push 2023-08-01), retrieved 2026-11-15.**
https://github.com/SupritYoung/RLHF-Label-Tool

18. **GitHub (5 stars, created 2026-09-25), retrieved 2026-11-15.**
https://github.com/Manirider/RLHF-Preference-Labeling-Platform

19. **GitHub (1,393 stars, last push 2026-09-29), retrieved 2026-11-15.**
https://github.com/xtreme1-io/xtreme1