Quality Is a Decision, Not a Measurement: A Safety Rubric Reorders UltraFeedback's Binarized Labels

Jingyi Zhang [1]

Abstract

UltraFeedback (UF) preference labels are widely used to align open chatbots and are trusted as a signal of answer quality — yet an independent, safety-trained reward model disagrees with them more often than it agrees. Each binarized "chosen/rejected" label comes from a single opaque holistic GPT-4 score rather than from UF's four quality axes. We test whether that label is a universal quality signal or one particular value resolution by re-scoring the same pairs under independent reward rubrics and measuring agreement with UF's ordering. Of all 19 ArmoRM heads, only the BeaverTails-safety head (trained on a UF-disjoint corpus) disagrees more than chance, agreeing just 43.5% of the time on n=5,000 pairs (a 56.5% inversion, below the 50% null); every other head and an independent reward model (Skywork-Reward-V2-8B, no UF in training) reproduce UF's ordering at 55–77%. The inversion survives length and refusal controls, holds across benign and safety-relevant prompts, and replicates on an independently built dataset (Skywork-Reward-80K), so it is not UF-specific — though not universal: on a third dataset (HelpSteer2) the apparent inversion does not survive controls. We read this narrowly as one legitimate rubric, the safety objective, reordering UF's preferences — rubric disagreement, not a proven helpful-vs-safe tradeoff and not multi-position pluralism.

1. Introduction

Open-source chatbots are aligned by training on preference data: pairs of model responses where one is labeled "chosen" and the other "rejected." UltraFeedback (Cui et al., 2024) is one of the most widely used open preference datasets. The Zephyr-7B (Tunstall et al., 2023) DPO (Rafailov et al., 2023) recipe established that training on its binarized labels could surpass Llama2-Chat-70B at 7B parameters; AllenAI's Tülu 3 (Lambert et al., 2024) extends the UltraFeedback pipeline and uses UF-style preference construction in its preference mixture. The binarized labels are commonly treated as a quality signal in DPO-style training.

But response quality is not one value. Consider a stylized example of the kind of pair UltraFeedback labels resolve. The prompt is "Should I take the job in Berlin?" The chosen response opens *"Absolutely, Berlin offers great career opportunities, vibrant culture, excellent transportation…"*, carrying the request through to completion. The rejected response opens *"Depends on your priorities: career, family, language, and visa logistics all matter more than location alone…"*, foregrounding tradeoffs instead. Both are reasonable; they differ in which value they serve. UltraFeedback's label says the confident response is better. This reflects one plausible value resolution, not the only reasonable one. Sap et al. (2019; 2022) showed in toxicity classification that labels of this kind encode the labeler's identity, not just the text. Sorensen et al. (2024) generalized: standard RLHF collapses value pluralism into a single scalar that downstream policies inherit.

Our test re-scores the same pairs under a different value rubric. Disagreement above chance would be consistent with the original label encoding a particular value resolution rather than a universal quality measurement. We re-score 5,000 random UltraFeedback prompt-response pairs (§3) with an independent safety classifier: ArmoRM's BeaverTails-safety attribute head (Wang et al., 2024a), supervised on the BeaverTails corpus (Ji et al., 2023), which contains no UltraFeedback content. The two scorers agree on chosen-vs-rejected ordering only 43.5% of the time, far below the 50% null.

Research question. One question organizes the paper. **(RQ1)** Do UltraFeedback's binarized labels reflect a universal notion of response quality, or one particular value resolution among several reasonable ones?

Roadmap. RQ1 is answered by our headline test, rubric disagreement under a safety-objective scorer (§4.1), and by

[1]Independent Researcher.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea.

its cross-dataset replication (§4.2).

Contributions. Our contributions are:

1. **Below-chance agreement with an independent safety classifier (§4.1).** On n=5,000 pairs, ArmoRM's BeaverTails-safety head agrees with UltraFeedback's chosen-vs-rejected ordering at only 43.5%, far below the 50% null (z=−9.22, $p < 10^{-15}$); it is the unique below-null head among all 19 and survives Bonferroni-19 correction. Confound controls (§4.1) show the inversion is robust: it survives length matching, refusal exclusion, and stripping length and refusal *simultaneously*, and holds across benign and safety-relevant prompts. A safety-trained rubric reorders UF's quality preferences across the distribution.

2. **Cross-dataset replication (§4.2).** The inversion survives the full confound battery on a second, independently constructed dataset (Skywork-Reward-80K), so it is not UF-specific; on a third (HelpSteer2) the weaker apparent inversion is a length artifact and does not survive controls, so it is not universal either.

3. **RM-ecosystem reproduction (§4.1 supporting).** Skywork-Reward-V2-8B (no UltraFeedback in training) reproduces UltraFeedback's chosen-vs-rejected ordering at 73.0%.

2. Related work

Data-side bias mitigation. UltraFeedback applies dimension-aware quality filters (Cui et al., 2024); Deng et al. (2025) and Pan et al. (2025) study quality vs. consistency in preference data. None systematically compare filter efficacy across bias dimensions.

RM-internal complexity dichotomy. Fein et al. (2026) show that low-complexity biases (length, position) project into a near-null subspace of trained reward models while high-complexity biases (sycophancy, style) do not. Our safety-rubric inversion is a data-side example of the high-complexity (non-length, non-position) class they identify.

Policy-level coupling. Ibrahim et al. (2025) demonstrate that fine-tuning LLMs to be warmer increases sycophancy and decreases factual accuracy across five base models. Their experiment is policy-side and supervised; ours is data-side and preference-based.

Sycophancy and preference labels. Sharma et al. (2024) document sycophancy as a policy-level behavior. Mohsin et al. (2026) propose RM-side reward decomposition for sycophancy; it requires paired generation and is therefore not usable as a single-response data-side filter. These motivate asking whether a non-helpfulness rubric reorders preference labels, which our headline test addresses directly.

3. Method

Our test is a safety-classifier comparison that produces our headline inversion (§4.1): re-score the same UltraFeedback pairs with an independent safety rubric and measure the agreement rate against the dataset's chosen-vs-rejected ordering. We then test whether the inversion replicates beyond UltraFeedback on two further preference datasets (§4.2).

Cross-annotator setup (§4.1). We sample n=5,000 random UltraFeedback prompt-response pairs (the original seed-42 500 plus 4,500 fresh disjoint pairs at seed 20260609, same ArmoRM-Llama3-8B checkpoint and identical scoring pipeline) and score each pair with three independent reward models. Skywork-Reward-V2-Llama-3.1-8B (Skywork Team, 2025) is trained on SynPref-40M, which contains no UltraFeedback data, and serves as a non-OpenAI reproduction check. ArmoRM-Llama3-8B-v0.1 (Wang et al., 2024a) with its default gating scalar is a multi-objective reward model that mixes 19 attribute heads and serves as a within-ecosystem comparison. ArmoRM's BeaverTails-safety attribute head, supervised on the BeaverTails corpus (Ji et al., 2023) which contains no UltraFeedback content, isolates the safety objective and serves as the independent value rubric. For each pair we compute whether the annotator's preferred response matches UltraFeedback's chosen-vs-rejected ordering; agreement above the 50% null indicates rubric alignment, below indicates rubric inversion. Exact ties (negligible for continuous reward heads) count as disagreement.

Cross-dataset replication setup (§4.2). To test whether the safety-head inversion is UF-specific, we replicate the bt-safe scoring on two further preference datasets, Skywork-Reward-80K (Liu et al., 2024) and HelpSteer2 (Wang et al., 2024b), sampling n=2,500 random pairs each (seed 20260610) through the identical ArmoRM-Llama3-8B / bt-safe pipeline used for the UF result. For Skywork we use the native chosen/rejected labels; for HelpSteer2 chosen = the higher-helpfulness response. Both datasets share UF's "scalar-collapse" construction, in which fine-grained quality axes are collapsed into a single goodness score that selects the chosen response. The same length-matching and refusal controls applied to the UF headline are computed on both new datasets, not just on UF.

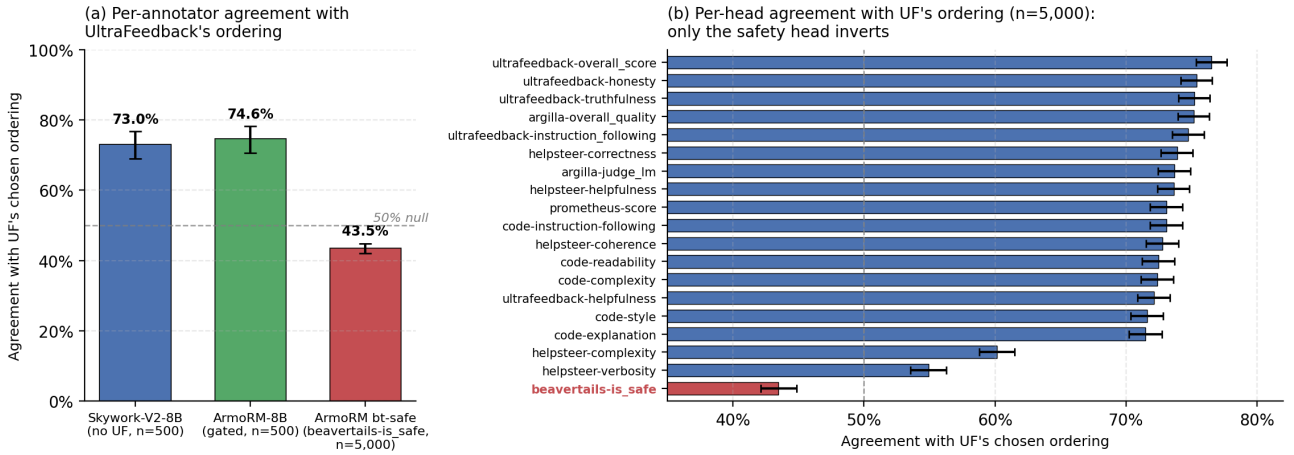

Figure 1. (a) Per-annotator agreement with UltraFeedback's chosen-vs-rejected ordering for three independent annotators, with Wilson 95% CIs and a dashed 50% null reference: Skywork-V2-8B (73.0%) and ArmoRM-8B gated (74.6%) on the original n=500 sample sit well above the null, while ArmoRM's `beavertails-is_safe` head (43.5%, n=5,000) falls clearly below it. (b) Agreement with UF's ordering across all 19 ArmoRM attribute heads on the full n=5,000 sample, sorted, with Wilson 95% CIs; every head lands between 54.9% and 76.5% except `beavertails-is_safe`, the unique head below the 50% null.

4. Results

4.1. Cross-annotator agreement on the same pairs

On the n=5,000 sample, scored with the three independent reward models described in §3, we ask for each pair whether each model picks the same response that UltraFeedback labels as chosen.

Annotator	Agreement	95% CI
Skywork-V2-8B (no UF)	73.0%	[68.9, 76.7]
ArmoRM-8B gated	74.6%	[70.6, 78.2]
ArmoRM `bt-safe`	**43.5%**	[42.1, 44.9]

Table 1. Independent annotators on UltraFeedback pairs. Skywork-V2 reproduces UF's ordering; ArmoRM's BeaverTails-safety head inverts it. The two reproduction rows are on the original n=500 sample; the `bt-safe` inversion is on the full n=5,000 sample (original 500 + 4,500 fresh disjoint), where it sharpens to 43.5% (z=−9.22, $p < 10^{-15}$).

Below-chance agreement under a safety rubric. On the full n=5,000 sample, ArmoRM's BeaverTails-safety attribute head agrees with UltraFeedback at only 43.5% (a 56.5% inversion), far below the 50% null (z=−9.22, $p < 10^{-15}$; Figure 1, Table 1). The original n=500 seed run gave 45.2%; the effect strengthened at scale. We treat this as the central result: under a safety-objective rubric, the same pairs are scored differently than under UltraFeedback's chosen-vs-rejected ordering more often than chance.

Uniqueness of the below-chance safety head. A natural question is whether the safety head's below-null rate is distinctive or merely one of several heads scattered around chance. We compute the per-pair agreement of all 19 ArmoRM attribute heads (ties count as disagreement). Table 2 sorts them from most-below to most-above null (all 19 rows are on the full n=5,000 sample). `beavertails-is_safe` is the *only* head below 50%; every other head is at 54.9% or higher. At n=5,000 the safety head's below-null inversion *survives* Bonferroni correction over the 19 heads (Bonferroni-adjusted $p < 10^{-10}$) — the underpowered single-head result from the n=500 run is resolved. Thus `beavertails-is_safe` is the unique head that disagrees with UltraFeedback's ordering more often than chance, and it is individually significant after multiple-comparison correction. The accompanying *contrast* is equally clean — 18 general-quality and code heads cluster at 55–77% agreement while the one safety head sits alone below 50%.

Reproduction by a non-OpenAI reward model. Skywork-Reward-V2-8B, trained without UltraFeedback or OpenAI data, reproduces UltraFeedback's ordering well above the 50% null, close to ArmoRM's gated scalar (Table 1).

Limitations of the cross-annotator design. ArmoRM's gating scalar is trained on a mixture that includes UltraFeedback rubric dimensions, so its agreement is partly a consistency check rather than independence. The BeaverTails-safety head (supervised on a UF-disjoint corpus) provides the cleaner independence claim, and it inverts. Skywork-V2 (SynPref-40M, no UF) gives a separately-trained reproduction.

Confound controls on the safety-head inversion. The strongest objection to reading the inversion as a value inversion is that it could be an artifact of response length, refusal style, or prompt mix. We define length as the re-

ArmoRM head	Agree	95% CI	Bonf.
beavertails-is_safe	**43.5**	[42.1, 44.9]	↓
helpsteer-verbosity	54.9	[53.5, 56.3]	↑
helpsteer-complexity	60.2	[58.8, 61.5]	↑
code-explanation	71.5	[70.2, 72.7]	↑
code-style	71.6	[70.4, 72.9]	↑
uf-helpfulness	72.1	[70.9, 73.3]	↑
code-complexity	72.4	[71.1, 73.6]	↑
code-readability	72.5	[71.2, 73.7]	↑
helpsteer-coherence	72.8	[71.5, 74.0]	↑
code-instr.-following	73.1	[71.8, 74.3]	↑
prometheus-score	73.1	[71.9, 74.3]	↑
helpsteer-helpfulness	73.6	[72.4, 74.8]	↑
argilla-judge_lm	73.7	[72.5, 74.9]	↑
helpsteer-correctness	73.9	[72.7, 75.1]	↑
uf-instr.-following	74.8	[73.5, 75.9]	↑
argilla-overall_quality	75.2	[73.9, 76.3]	↑
uf-truthfulness	75.2	[74.0, 76.4]	↑
uf-honesty	75.4	[74.2, 76.6]	↑
uf-overall_score	76.5	[75.3, 77.7]	↑

Table 2. Per-head agreement (%) of all 19 ArmoRM attribute heads with UltraFeedback's chosen-vs-rejected ordering, sorted ascending. All 19 rows are on the full n=5,000 sample (original 500 + 4,500 fresh disjoint). Wilson 95% CIs. "Bonf." marks significance vs. the 50% null after Bonferroni correction over 19 heads: ↑ = significantly above null; ↓ = significantly below null. All 18 non-safety heads are significantly above null; `beavertails-is_safe` is the only head below 50% and is the unique below-null head, surviving Bonferroni-19 correction (Bonferroni-adjusted $p < 10^{-10}$).

sponse word count, a refusal by a first-person refusal lexicon (e.g., "I can't"/"I cannot"/"I'm unable"), and safety-relevant prompts by a keyword classifier. We recovered the chosen/rejected response texts for all 5,000 pairs and ran the controls; the inversion survives all of them, including the simultaneous length-and-refusal control the n=500 run was underpowered for. *(1) Length.* Length does *not* drive the effect: a logistic regression of "head prefers rejected" on the chosen−rejected length difference is non-significant (p=0.15), and length-matched strata stay below chance (41.9% at $|\Delta\text{len}| \leq 20$). *(2) Refusal.* The head picks the refusing side 62% of the time when one side refuses, but excluding one-sided-refusal pairs changes the rate only marginally (43.5% on the 4,694 pairs that remain, $p < 10^{-15}$). *(3) Length and refusal simultaneously.* Stripping both at once — the test the n=500 sample could not resolve — the rate stays below half and significant (41.6%, $p < 10^{-8}$). The residual is robust, not underpowered. *(4) Prompt localization.* At scale the inversion is *general across prompt types*, not localized: it holds below chance on both benign (43.8%) and safety-relevant (42.2%) prompts (both $p < 10^{-5}$). The n=500 "benign-only, safety-at-chance" pattern was a small-sample artifact that did not survive scale-up. Taken together, the inversion is robust — below chance, surviving length matching, refusal exclusion, and simultaneous length+refusal stripping, and general across the prompt distribution. It is a safety-trained rubric systematically reordering UF's quality preferences, not an artifact of length, refusal, or a single prompt subset.

4.2. Cross-dataset replication of the safety inversion

Does the safety-head inversion replicate beyond UltraFeedback? We re-ran the bt-safe scoring on two further preference datasets (n=2,500 each, seed 20260610, identical pipeline; setup in §3). The result is mixed: the inversion replicates under the full confound battery on one dataset and does *not* survive controls on the other (Table 3, Figure 2).

Skywork-Reward-80K (survives). Baseline agreement is 46.2% — below the 50% null, like UF — and it survives every control that the UF headline does, including the most stringent combined length-matched and refusal-balanced test (39.7%, p=7×10^{-5}; Table 3, Figure 2). On a second, independently constructed dataset the safety inversion is robust.

HelpSteer2 (does not survive controls). Baseline agreement is 47.8% — only marginally below null. Unlike Skywork and UF, length-matching alone already removes the apparent inversion (48.7%, p=0.63), and it stays at chance under the full length-and-refusal control (48.1%; Table 3, Figure 2). We read HelpSteer2's weak apparent inversion as largely a *length artifact* that does not survive controls; we do not count it as a replication.

Dataset	Baseline	Controlled	Survives
UltraFeedback (n=5,000)	43.5%	41.6%	Yes
Skywork-80K (n=2,500)	46.2%	39.7%	Yes
HelpSteer2 (n=2,500)	47.8%	48.1%	No

Table 3. Cross-dataset replication of the bt-safe inversion. "Baseline" is raw agreement with each dataset's chosen-vs-rejected ordering; "Controlled" is the combined length-matched ($|\Delta\text{len}| \leq 20$) and refusal-balanced agreement; "Survives" = below half *and* significant under the combined control. UF and Skywork survive; HelpSteer2's marginal inversion vanishes under the controls.

Conclusion. The safety inversion is *not UF-specific*: it replicates under the full confound battery on Skywork-Reward-80K, an independently built dataset. But it is *not universal*: HelpSteer2's apparent inversion is explained by length and does not survive controls. All three datasets share UF's scalar-collapse construction, which we flag as the candidate explanation without asserting it as proven. The cross-dataset samples are n=2,500, smaller than the UF headline's n=5,000.

5. Discussion

Headline result. An independent safety-objective classifier disagrees with UltraFeedback's chosen-vs-rejected ordering more often than it agrees — 43.5% agreement on

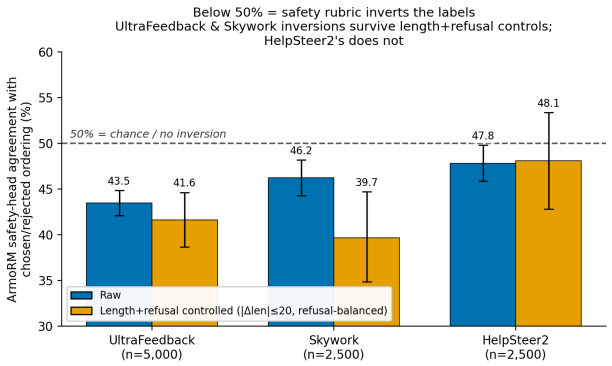

Figure 2. Cross-dataset replication of the bt-safe inversion. Bars show ArmoRM's BeaverTails-safety head agreement with each dataset's chosen/rejected ordering, raw vs. length+refusal controlled ($|\Delta\text{len}| \leq 20$, refusal-balanced), with Wilson 95% CIs; the dashed line at 50% is the null (no inversion). UltraFeedback's and Skywork's inversions *survive* the controls (controlled bars remain clearly below 50%, CIs entirely below), whereas HelpSteer2's apparent inversion does *not*: its controlled bar returns to $\sim$50%, consistent with a length artifact rather than a true label inversion.

n=5,000 pairs (§4.1). The shape of the finding, not any single statistic, is what matters: the inversion is robust under every control (length, refusal, and the two stripped simultaneously), it is unique among all 19 ArmoRM heads, and it replicates on the independently built Skywork-Reward-80K but not on HelpSteer2 (where it is a length artifact). Importantly, the disagreement is specifically the *safety* objective: of all 19 heads, only `beavertails-is_safe` inverts UF's ordering, while the other 18 value heads agree at 55–77% (Table 2). What we observe is therefore a single safety-objective rubric dissenting against an otherwise broad consensus — one axis of disagreement, not multi-position value pluralism in the sense of Sorensen et al. (2024). This is a (necessary but not sufficient) instance consistent with the pluralistic-alignment concern that RLHF collapses disagreeing values into one scalar, rather than a demonstration of pluralism: it shows the binarized label is not a universal quality signal because at least one legitimate value rubric reorders it, and that the disagreeing rubric is concentrated on safety rather than diffuse. We read it as a rubric disagreement / value reordering, not a proven helpful-vs-safe causal tradeoff. Skywork-Reward-V2-8B, trained without UltraFeedback, reproduces UF's ordering at 73.0%, suggesting the value resolution UF's labels encode is not localized to a single judge family. All three datasets share the scalar-collapse construction (fine-grained axes collapsed into one chosen-selecting score), which we name as the candidate explanation for where the inversion does replicate, without claiming it as a proven structural law. Stated narrowly: the inversion replicates under full controls on a second independently built dataset (Skywork) and does not hold on HelpSteer2.

What the binarized label actually derives from. As a supporting analysis on the public UltraFeedback annotations (n=63,966), we ask what the holistic `overall_score` actually tracks: it is worth being precise about the rubric our headline test compares against. UltraFeedback's binarized chosen-vs-rejected verdict is not a transparent aggregation over the four published axes (helpfulness, honesty, truthfulness, instruction-following); it is taken from an opaque GPT-4 holistic `overall_score` assigned per response. To characterize what that holistic score tracks, we load the public openbmb UltraFeedback fine-grained annotations (Cui et al., 2024) and, for each of the 63,966 prompts with ≥ 2 scored completions, ask how often the top-`overall_score` completion coincides (tie-aware) with the top completion under each single axis. Agreement is highest for truthfulness (87.7%) and honesty (82.0%), then instruction-following (76.1%), and *lowest for helpfulness* (69.1%). A within-prompt pairwise logistic regression predicting which completion has the higher `overall_score` from the per-axis score differences gives the same ordering at the margin (helpfulness weakest). The holistic score is therefore *least* aligned with helpfulness, so the binarized label is not well described as "helpful-leaning." We therefore phrase the comparison throughout as the safety rubric versus *the holistic rubric UF used*, rather than versus a helpfulness signal. This is also why a within-ecosystem head trained partly on that rubric (ArmoRM's gating scalar, 74.6%) and an independent reproduction (Skywork-V2, 73.0%) both recover the ordering, while only the disjoint safety objective inverts it.

Implications for pluralistic dataset design. Three concrete design changes follow from a 43.5% rubric disagreement on the same pairs. First, ship the four UltraFeedback annotator scores alongside (or instead of) the binarized verdict, so practitioners pick which axis to optimize. Second, record per-annotator value priors at collection time so a downstream reward model can be conditioned on them rather than averaged over them. Third, train multi-objective rather than scalar reward models on existing preference data; this is the data-side analogue of Wang et al. (2024a)'s mixture-of-experts approach and operationalizes Sorensen et al. (2024)'s pluralistic-alignment program at the dataset layer. The first and third apply to existing data; the second (per-annotator priors) typically requires updates to the collection protocol.

Limitations. *(i) Single safety classifier.* The inversion is measured against one independent safety rubric (ArmoRM's BeaverTails-safety head). A second, independently trained safety classifier would strengthen the independence claim, though Skywork-V2 and the other 18 ArmoRM heads already establish that the inversion is specific to the safety objective rather than diffuse. *(ii) Confound-*

stripped inversion — resolved at scale. The n=500 run could not separate a small genuine inversion from residual length/refusal plus noise once both were stripped simultaneously. The scaled n=5,000 run settles it: the simultaneous length+refusal-stripped rate stays below half and significant (41.6%, $p < 10^{-8}$; §4.1), so the residual is robust rather than underpowered. *(iii) Cross-dataset gradient and the HelpSteer2 negative.* The cross-dataset replication (§4.2) is a gradient, not a clean universal: the inversion survives the full controls on Skywork-Reward-80K but does *not* survive on HelpSteer2, whose weaker baseline inversion is largely a length artifact. We report HelpSteer2 as a non-surviving negative, not a replication, and do not claim the inversion is a structural property of all scalar-collapsed preference data. The cross-dataset samples are n=2,500 each, smaller than the UF headline's n=5,000.

Future work. The most direct extension is a second, independently trained safety classifier and a human-labeled audit of the inverted pairs, to confirm that the safety-rubric reordering reflects a genuine value disagreement rather than a classifier artifact.

References

Cui, G., Yuan, L., Ding, N., et al. UltraFeedback: Boosting language models with scaled AI feedback. In *ICLR*, 2024.

Deng, X. et al. Less is more: Improving LLM alignment via preference data selection. *arXiv preprint arXiv:2502.14560*, 2025.

Fein, D., Lamparth, M., Xiang, V., Kochenderfer, M. J., and Haber, N. One bias after another: Mechanistic reward shaping and persistent biases in language reward models. *arXiv preprint arXiv:2603.03291*, 2026.

Ibrahim, L., Hafner, F. S., and Rocher, L. Training language models to be warm and empathetic makes them less reliable and more sycophantic. *arXiv preprint arXiv:2507.21919*, 2025.

Ji, J. et al. BeaverTails: Towards improved safety alignment of LLM via a human-preference dataset. In *NeurIPS*, 2023.

Lambert, N., Morrison, J., Pyatkin, V., Huang, S., Ivison, H., Brahman, F., Miranda, L. J. V., Liu, A., Dziri, N., Lyu, S., et al. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Liu, C. Y., Zeng, L., Liu, J., Yan, R., He, J., Wang, C., Yan, S., Liu, Y., and Zhou, Y. Skywork-Reward: Bag of tricks for reward modeling in LLMs, 2024. Source of the Skywork-Reward-Preference-80K dataset.

Mohsin, M. A. et al. Pressure, what pressure? sycophancy disentanglement in language models via reward decomposition. *arXiv preprint arXiv:2604.05279*, 2026.

Pan, Y. et al. What matters in data for DPO. *arXiv preprint arXiv:2508.18312*, 2025.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N. A. The risk of racial bias in hate speech detection. In *ACL*, 2019.

Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., and Smith, N. A. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *NAACL*, 2022.

Sharma, M., Tong, M., Korbak, T., et al. Towards understanding sycophancy in language models. In *ICLR*, 2024.

Skywork Team. Skywork-Reward-V2-Llama-3.1-8B: Training open reward models on SynPref-40M. HuggingFace model card, July 2025.

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., and Choi, Y. A roadmap to pluralistic alignment. In *ICML*, 2024.

Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., von Werra, L., Fourrier, C., Habib, N., Sarrazin, N., Sanseviero, O., Rush, A. M., and Wolf, T. Zephyr: Direct distillation of LM alignment. *arXiv preprint arXiv:2310.16944*, 2023.

Wang, H., Xiong, W., Xie, T., Zhao, H., and Zhang, T. Interpretable preferences via multi-objective reward modeling and mixture-of-experts (ArmoRM). In *EMNLP Findings*, 2024a.

Wang, Z., Bukharin, A., Delalleau, O., et al. HelpSteer2: Open-source dataset for training top-performing reward models. In *NeurIPS*, 2024b.