(IN)TRANSITIVITY: WHERE BRADLEY-TERRY FAILS TO MODEL HUMAN PREFERENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Most preference-based alignment methods use Bradley-Terry (BT) preference models, which implicitly assume human preferences are consistent and transitive, i.e., that they can be collapsed onto a single scalar. In reality, however, heterogeneous human preferences can cause intransitivity through cyclic majority relations both at the user, as well as population level—a problem well-studied in social choice theory. While recent game-theoretic alignment methods propose cycle-aware objectives to handle this, whether such cycles matter empirically remains unclear. In this work, we answer this question by first auditing 9 widely used human-preference datasets and find that by construction, most lack the overlapping, repeated judgements required to analyze population-level intransitivity. We then construct INTRANSIT, a controlled dense preference dataset, through preference-completion methods that estimate the missing population-level preferences. We use this as a measurement tool for widely used preference models and show how performance degrades to near-chance agreement on intransitive samples. At the same time, the denser preference structure in the completed datasets contains useful signal in that preference models trained on this data perform as well as models trained on fully complete pairwise annotations. Finally, at inference time, using trained preference model outputs reduces regret and majority-rejected selections, and reinforcement-learning systems trained with preference models over this show improvements for certain preference objectives. Our results identify two challenges for preference alignment: collecting and measuring preferences to allow intransitive preferences to surface, as well as building objectives for the representation of these preferences without imposing a scalar ordering.

1 INTRODUCTION

Language model post-training predominantly relies on human preference judgements to align with humans. In reinforcement learning from human feedback (RLHF), human annotators compare candidate responses, and a dataset of such preference comparisons is used to learn a reward function that ranks responses in order of preference Stiennon et al. (2020). A standard formulation is the Bradley-Terry (BT) model that represents each response with a reward thus allowing for the modeling of pairwise preferences as a function of reward differences Bradley & Terry (1952); Ouyang et al. (2022). By construction, this model makes a consequential assumption: that the preferences being modeled follow a scalar and transitive ordering. Therefore, it assumes that if a response A is preferred to B and B to C, then A is preferred to C.

However, human preferences need not have this structure. Irrationality in human decision-making has long been studied in individual choice (Tversky, 1969; Regenwetter et al., 2011), where a single user can exhibit choice cycles in the presence of multiple conflicting attributes. At the population level, social choice theory has long established that cyclic majority relations can exist even when individual users exhibit transitivity (Arrow, 1951; Sen, 1966). Even in principle, a BT model cannot represent a violation of the transitive assumption leading to a consequence for language-model alignment: when models encounter a cycle, a scalar model must discard at least one relation in the cyclic graph. This concern has motivated a recent line of work proposing alignment methods that do not collapse preferences to scalar rewards, but are more *cycle-aware* through non-scalar objectives (Munos et al., 2023; Swamy et al., 2024; Rosset et al., 2024). However, an empirical question under-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

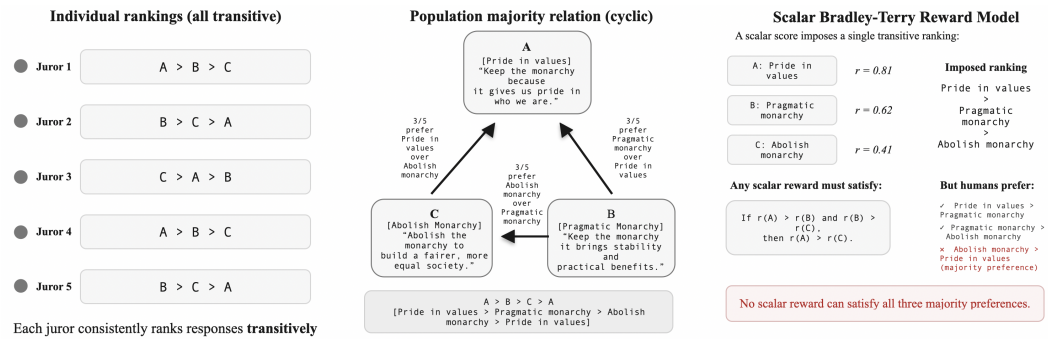


Figure 1: **Human judgements can produce cyclic population-level rankings** Five jurors from the Habermas study (Tessler et al., 2024) each rank three samples such that their majorities form a cycle $(A \succ B, B \succ C, C \succ A)$. A scalar BT reward model on the right cannot correctly represent this.

lying these works remains unanswered, which is, do human preference datasets contain intransitive cycles? And if so, does this affect downstream language-model alignment performance?

We answer this question in three stages. First, in §2 we ask whether existing preference datasets can reveal intransitivity in their annotations. We audit 9 widely-used preference datasets and find that most are structurally unable to reveal cycles because they only satisfy disjoint sets of properties required to identify majority cycles. As a result, the apparent absence of reported cycles reflects a dataset annotation limitation rather than evidence of transitivity of human preferences. Second, we ask whether the missing preference structure can be recovered by "completing" these datasets in §3. We develop INTRANSIT, a dataset that serves as a controlled dense preference benchmark grounded in 1.3k real annotator profiles, candidate responses and human judgements, and develop a population-level preference completion method to predict missing pairwise relations. We validate several completion methods on held-out human judgements and use the best-performing method that recovers unseen preferences well but tends to underestimate contested cyclic data, to construct the final dataset. Lastly, we evaluate the effect of these completed preferences on downstream alignment methods in §4. We fit BT models to different preference datasets, and show that the completed preferences provide useful training signal: preference reward models trained on completed datasets approach performance of exhaustive pairwise annotations. At inference-time, we see that using the completed preference predictions reduces both regret and majority-rejected selections. We also see mixed gains on post-training with this data; where some preference training objectives (e.g., DPO (Rafailov et al., 2023)) show improved performance, but others (IPO, KTO) show no improvements.

Put together, our results identify two important problems for preference alignment: measurement, since existing data collection regimes do not allow for intransitive preferences, as well as modeling, since preference objectives enforce scalar ordering. We show benefits of considering intransitive preferences that imply that alignment pipelines should test this assumption as well as preserve richer data when possible. We summarize our contributions as follows:

1. **A data-driven audit of intransitivity in preference datasets.** We evaluate several publicly available preference datasets for criteria required to identify population cycles and see that most, by construction, do not contain enough per-sample or per-rater annotations to reveal them.

2. **A methodology to recover missing preference structure.** We develop and validate preference-completion methods and use this to construct INTRANSIT. This lets us estimate missing preference comparisons to augment sparse preference datasets.

3. **Measuring model performance on intransitive data.** We evaluate several predominantly used preference models on completed datasets and see that models degrade sharply on cyclic samples, demonstrating the practical consequences of scalar preference reward structures.

4. **Using completed preference information downstream** We see that completed preferences improve training of preference reward models, have small post-training benefits, and improve inference-time serving to reduce regret and majority-rejected selections.

2 Does Intransitivity Exist in Preference Datasets?

2.1 Preference Datasets

We audit 9 preference datasets used in language-model alignment. These datasets vary in how preferences are collected in several ways: the number of raters that judge each sample (single vs. repeated judgements), preservation of anonymized annotator idenities (to track per-annotater judgements across samples), collecting pairwise judgements vs. picking one out of four samples and so on. We briefly outline the datasets below, and then compute properties of datasets and cycle existence across datasets in §2.2 and §2.3.

The datasets we consider fall into the following annotation structures: **Best-of-(k) datasets** include PRISM (Kirk et al., 2024), where participants evaluate up to four model responses to their own prompts and feedback is linked to persistent user profiles, and Community Alignment (CALign) (Zhang et al., 2026), which similarly elicits a choices from sets of four. **Pairwise datasets** include the Summarization-from-Human-Feedback (SHF) dataset (Stiennon et al., 2020), Anthropic HH-RLHF (Bai et al., 2022), MultiPref (Miranda et al., 2025), and LMSYS (Chiang et al., 2024), which provide comparisons between two responses at a time and therefore require repeated, overlapping comparisons to form observable cycles. HelpSteer2 (Wang et al., 2024) instead provides fine-grained attribute ratings of responses but lacks pairwise comparisons entirely. Finally, we consider two datasets that contain substantially denser population feedback: the Habermas dataset (Tessler et al., 2024) that provides full rankings from small groups deliberating over candidate statements, and OvertonBench (Poole-Dayan et al., 2026) that collects repeated 1–5 representation ratings from large participant samples across responses to value-laden questions. These latter datasets are smaller than the former, but provide unusually direct views of population-level preference structure. We also considered other datasets with multiple annotators however some, such as DICES (Aroyo et al., 2023; Rastogi et al., 2026) contain safety ratings rather than preferences, and others such as Stanford Human Preferences (SHP) Ethayarajh et al. (2022) that are structurally similar to the datasets above and do not add new audit categories.

Dataset	Unique Raters	Raters / Sample	Total Samples	Annotation Format	Overlapping Comparisons	User IDs	Repeated Judgements	Cycles Possible	Cycle Rate
PRISM	1,309	1	8,011	choose 1 of 4	✓	✓	×	×	-
SHF	76	1	178,944	pairwise	✓	✓	≈	×	-
MultiPref	pseud.	2	5,323	pairwise	×	✓	✓	×	-
HelpSteer2	none	1	20,324	attribute ratings	×	×	×	×	-
CAlign	3,603	1	33,099	choose 1 of 4	≈	✓	≈	×	-
HH-RLHF	none	1	169,352	pairwise	×	×	×	×	-
LMSYS	pseud.	1	33,000	pairwise	≈	✓	×	≈	-
Habermas	5,734	5	7,340	full ranking	✓	✓	✓	✓	4.7%
OvertonBench	1,130	59	60	1-5 rating	✓	✓	✓	✓	8.3%
InTransit	1,379	1,379	3,700	derived ranking	✓	✓	✓	✓	9.4%

Table 1: **Auditing Preference Datasets.** 9 preference datasets are audited for annotation structure required to observe cyclic majority relations. Only Habermas and OvertonBench provide repeated, overlapping, and user-preserving judgements that allow for cycle computation. We compare this to InTransit that we construct in §3 and the properties it shows.

2.2 Detecting Cycles in Datasets

We represent the preferences associated with a prompt as a directed graph ($G = (V, E)$), where each node ($y_i \in V$) is a candidate response and an edge ($y_i \to y_j$) indicates that (y_i) is preferred to (y_j). We focus on the smallest such intransitive structure, a three-response cycle, ($y_i \succ y_j, y_j \succ y_k, y_k \succ y_i$) that cannot be represented by any scalar ordering over the three responses. We separate

two sources of cyclic structure: *individual intransitivity*, where one annotator's own judgments form a cycle, and *population intransitivity*, where the majorities form a cycle even though every annotator is transitive, as in the Condorcet paradox (Arrow, 1951; Sen, 1966). We focus on the latter.

In the constructed graph, we call a response triple an *observable cycle* if all three pairwise relations exist over the population. We treat missing edges as unobserved rather than transitive. For a given dataset, we therefore enumerate all observable triples and test whether their directed majority relation forms a cycle. We report this cycle rate as the fraction of eligible triples that are cyclic. When several annotators judge the same pair, we quantify uncertainty in each population relation through resampling, allowing us to distinguish cycles supported by stable majority preferences from those that may arise from finite-sample variation (see Appendix A.1 for resampling experiments and Appendix F.2 for further robustness checks).

2.3 Auditing Preference Datasets

In Table 1, we audit public preference datasets for the above criteria and reveal a big measurement limitation—that most datasets, given their annotation structures, cannot be evaluated for population cycles. Some datasets (e.g., PRISM) only contain one human judgement per response pair, while others (e.g., CommAlign) contain multiple judgements but lack repeated judgements from the same rater or sufficiently overlapping comparisons. Moreover, where some datasets retain anonymized annotator identification, most of them strip this information before release, making individual-level intransitivity for a rater impossible to distinguish from aggregation effects. We note two datasets that do contain dense human feedback, either through repeated or multiple annotators per sample: Habermas and OvertonBench containing a cycle rate of 4.7 and 8.3 respectively. This establishes that cycles can exist, when annotation structures allow them to be seen. However, these datasets were constructed to elicit behaviors of deliberation and consensus, thus their cycle rates should not be interpreted as estimates of naturally occurring intransitivity, but only as evidence that intransitivity is detectable when the data supports it.

Our audit reveals a problem of measurement: that existing preference datasets do not contain overlapping or repeated judgments to determine whether their preferences are cyclic. This motivates our question in §3 i.e., can the missing preferences in sparse datasets be estimated in order to "complete" these datasets and evaluate intransitivity?

3 Completing Sparse Preference Datasets

3.1 Preference Completion Methods

Although the measurement problem we identify can be corrected with new data collection, we acknowledge that exhaustively collecting the type of data required is expensive. An exhaustive collection over n candidate responses requires $\binom{n}{2}$ total pairs, each ideally rated by multiple annotators. In this section we therefore ask the following question: can the missing structure be recovered by estimating the relations that don't exist in current sparse preference datasets? We formulate this as a *preference completion* task. For a dataset with observed preferences (E_{obs}) over response pairs, we aim to fill in the missing preferences for every unobserved pair (y_i, y_j) by predicting the fraction of the population that prefers that ordering $[p_{ij} = P(y_i \succ y_j)]$. This is rooted in work on general preference modeling (Zhang et al., 2024b) and gives us a more robust target: rather than predicting the winning response, we predict the strength of the population disagreement itself.

Creating dense preference graphs for InTransit To evaluate this methodology we require held-out dense preferences against which estimated preferences can be compared. We construct InTransit from 1,379 annotators profiles in PRISM, choosing this dataset because it records annotator's relative assessments over three response attributes we focus on—helpfulness, safety and fluency. Therefore, an annotator's attribute ordering over new responses can be estimated. However, the responses in this dataset are sparse, so we use PersonalLLM which consists of prompts and responses from eight frontier LLMs. We measure each candidate response along the response attributes and use annotator's observed priority order to estimate annotator-specific rankings. We then use rank-order centroid weights (Barron & Barrett, 1996) to convert each priority ordering into weights resulting in a ballot i.e., a complete response ranking for each annotator over candidate

Table 2: **Completion Model Performance.** We report results on accuracy, error and cycle-accuracy on held-out human preferences across three datasets. GPM-Pop more accurately predicts preferences on contested subsets by a wide margin, but its calibration error remains imperfect, and no method can recover cycles in Habermas.

Dataset	Method	Acc(Decisive) ↑	Acc(Contested) ↑	MAE ↓	ECE ↓	Cycles
INTRANSIT	GPM-Pop	**0.772**	**0.582**	0.239	0.167	**17/22**
	PAL	0.771	0.485	0.237	0.102	9/22
	VPL	0.765	0.522	0.251	0.125	14/22
	Per-Fit	0.766	0.507	**0.108**	0.141	3/22
	GPM-Gen	0.658	0.425	0.135	**0.029**	6/22
Habermas	GPM-Pop	**0.693**	**0.543**	0.235	0.149	0/42
	Per-Fit	0.559	0.518	**0.101**	**0.039**	0/42
CAlign.	GPM-Pop	**0.757**	**0.641**	0.104	**0.021**	-
	Per-Fit	0.713	0.606	**0.067**	0.074	-

responses. See Appendix A.1 for details of construction and validation; one annotator's ballot is worked through in Table 6. We also require cycles to survive resampling, specifically, two bootstraps over both the annotator population as well as the response-attribute measurements. The resulting dataset contains 19 certified majority cycles over 14 prompts and 26,201 annotator ballots. Notably, none of the 26,201 ballots is itself intransitive; hence showing how the cycles arise from aggregating heterogeneous but individually transitive preferences.

Estimating missing preferences for population-level completion One approach to complete missing preferences is to train a user simulator for each annotator Wang et al. (2025). However, Table1 shows how most preference datasets do not contain enough repeated judgements to do this, therefore we turn to modeling the population directly. We build on GPM (Zhang et al., 2024b), a model that learns K-dimensional representations of responses according to latent population components. For a response pair (y_i, y_j), the model then predicts the population fraction $P(y_i \succ y_j)$ as the mixture of weighted components that prefer the first response. We adapt this into a population-based GPM (GPM-Pop) that predicts the share of a population preferring each response. On INTRANSIT, the population is the 1,379 annotators, each with their recorded weights; on public datasets without such records, the model learns a small number of latent groups. See Appendix C.3 for details of both. When fine-tuning on INTRANSIT we replace GPM's scalar pair score with two heads: one for K attributes that a response might have, and another for B groups that the population can be split into. The predicted fraction here therefore tells us about the share of the population whose group prefers one response over the other. During training, we upweight contested pairs three- to four-fold, since they are rare, and to deliberately optimize the model towards the contested regions that are sparse in human data. We report architecture, training, and dataset details for these models in Appendix C and evaluate performance against baselines that we describe in the next section.

3.2 EVALUATING COMPLETION METHODS

Data We evaluate the different completion methods on their ability to accurately predict human preferences on a held-out dataset. We hold out 28 prompts from INTRANSIT, each with all eight of its responses, yielding 784 response pairs and 1,568 triplets. For each pair, we compare the predicted population fraction with the ground truth, and check whether the method predicts the response that most of the 1,379 annotators prefer. We distinguish **decisive pairs**, where at least 65% of annotators preferred one response, from **contested pairs** where annotator division is split between 35% and 65%. We hypothesize that the latter is particularly important for intransitivity since cycles might arise at regions with substantial disagreement.

Metrics We evaluate completion performance at both the pairwise and graph levels with the following metrics. For held-out pairs, we compute majority accuracy (**Acc**) that tells us whether the model correctly predicted the response preferred by more than half the population and we report this on the **decisive** and **contested** pairs separately. We also report mean absolute error (**MAE**) between the predicted and ground-truth population preference fractions and calibration error (**ECE**). Over

the completed graph structures we evaluate the fraction of true cycles recovered among the triples that form real cycles in the dataset (**Cycles**) that tells us the precision of intransitive cycles being recovered.

Models We compare our population-based GPM (**GPM-Pop**) against two reward models that learn a personalized preference for each annotator. The Pluralistic Alignment model (**PAL**) (Chen et al., 2025) learns a shared set of reward functions and assigns each annotator individual non-negative weights. The Variational Preference Learning model (**VPL**) (Poddar et al., 2024) encodes each annotator's recorded preferences into a latent vector, which is then used as an input to the reward model. We also fit a ridge regression on response embeddings to each annotator's 25 comparisons (**Per-Fit**) and average the 1,379 fitted models into a predicted fraction per pair. We use the same baseline on Habermas and Community Alignment. Finally, we compare the general pre-trained GPM model (**GPM-General**). PAL, VPL and Per-Fit use 25 labels per annotator; GPM-Pop and GPM-General do not use any. See Appendix C for training configurations of these models.

Table 2 shows the difference in models performance on different parts of the dataset. On INTRANSIT all models except GPM-General perform similarly on decisive pairs reaching 77% majority-direction accuracy. On contested pairs, we see significant gains with GPM-Pop achieving a 58.2% accuracy compared to 48.5% and 52.2% for PAL and VPL, respectively. We also report the cycle-prediction accuracy on intransitive triples where GPM-Pop correctly predicts more cycles than other methods. Per-Fit and GPM-General have the lowest MAE on contested pairs simply because they predict fractions near 0.5—the same conservatism that reduces their accuracy. Our completion methods are therefore not to be interpreted as equivalent to individual human annotations, but rather an estimate of population-level preference structures; where the largest remaining errors are concentrated where the population is most divided. We also compare two other dense datasets and see similar trends: GPM-Pop outperforms Per-Fit on both decisive and contested samples, with the larger gain on decisive ones. On Habermas however, the GPM-Pop model fails to recover any of the true cycles in the dataset. Overall, this shows how population-level completion achieves sufficient accuracy on completion, and is able to recover harder, contested samples that contribute to dense preference structure.

3.3 PROPERTIES OF COMPLETED DATASETS

Pairwise prediction accuracy, even if precise, does not tell us about structural differences in the completed datasets. We therefore compare graph-based properties of these datasets, along with distributional analyses of preferences that are estimated. This tells us whether completion methods that fill in missing edges significantly change properties of the datasets in comparison to the original human annotated versions. We consider 1,200 prompts with dense population graphs, of which 24.8% contain a cycle. When GPM-Pop predicts every pair of the same prompts from the response text, only 11.9% contain a cycle. Thus, completion makes populations look more transitive without softening how divided they are—the mean majority margin moves from 0.394 in the real data to 0.391 after completion. We are still able to recover intransitive samples in commonly contested cases. This distinction matters: completion methods underestimate intransitivity precisely for the reason that missing preferences hardest to predict are the contested preferences likely to cycle. Especially in contested regions, rather than a replacement of direct human measurement, we view this as a method for augmenting sparse data and demonstrate the usefulness of this downstream in the next section.

We also compare the distribution of the preference landscape before and after completion adapting the dataset cartography approach from Swayamdipta et al. (2020). For response

Figure 2: **Dataset map** for original (top) and completed (bottom) preferences. Points represent unseen response pairs, blue/red indicate contested/decisive preference sets. Completed preferences shift models towards towards high confidence, but contested relations are still rated as high variability.

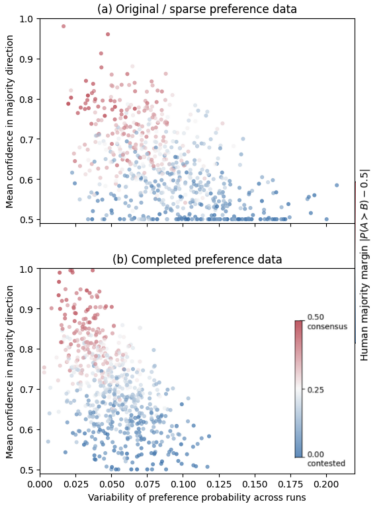


Table 3: **Preference models fail on intransitive samples, but training on completed data improves overall accuracy.** Comparing three model families on sparse vs. completed datasets shows improved overall accuracy, regret and transitive agreement. All scalar models degrade on intransitive subsets. A scalar ordering can satisfy at most two of three relations in a cycle (0.67) agreement.

		INTRANSIT				Habermas			
Objective	**Trained on**	**Acc. ↑**	**Trans./Intrans.**	**Regret ↓**	**ECE ↓**	**Acc. ↑**	**Trans./Intrans.**	**Regret ↓**	**ECE ↓**
Bradley-Terry	sparse	0.671	0.673 / 0.529	0.307	0.149	0.661	0.662 / 0.542	0.368	**0.013**
Bradley-Terry	completed	**0.719**	0.721 / 0.503	**0.272**	**0.047**	**0.688**	0.694 / 0.519	**0.349**	0.090
IPO	sparse	0.628	0.630 / 0.509	0.350	**0.036**	0.644	0.644 / 0.528	0.378	**0.102**
IPO	completed	**0.705**	0.706 / 0.523	**0.289**	0.070	**0.671**	0.675 / 0.528	**0.364**	0.145
Rank-4 head	sparse	-	0.651 / 0.513	0.299	-	0.650	0.651 / 0.532	0.370	**0.067**
Rank-4 head	completed	-	0.701 / 0.539	**0.291**	-	**0.682**	0.688 / 0.528	**0.352**	0.135

pairs in the datasets, we measure the majority margin ($m_{ij} = |P(y_i \succ y_j) - 0.5|$) where larger values indicate annotator consensus and lower values close to 0 indicate disagreement. This tells us about the samples that fall into contested regions that models are uncertain about in the different datasets. We color each data sample accordingly to this difficulty estimate in Figure 2. We see that completion shifts the preference landscape towards lower variability and higher confidence, whereas also increasing coverage over contested regions. Together, with the reduced cycle rate above, this suggests a limitation for future work: although completion is expected to model the original data that does not have intransitivity, it does not cover the parts of the map that have contested judgements—which are the ones most relevant to intransitivity.

4 USING COMPLETED PREFERENCES DOWNSTREAM FOR ALIGNMENT

4.1 HOW DO SCALAR PREFERENCE MODELS PERFORM ON CYCLES?

In this section, we validate the theoretical assumption we make about BT preference models: that their scalar reward modeling criteria does not allow handling of cyclic preferences. We compare three preference objectives: a BT model, IPO and a Rank-4 head objective.

All three use the same response embeddings. Both BT and IPO assign each response a single score but differ in their loss functions. BT directly fits the observed preference probability whereas IPO regresses the score difference to a predefined margin target. The Rank-4 head instead scores each response along four dimensions, with separate weighting for each value bloc. See Appendix D.1 for details.

We hold out 300 of 1,200 INTRANSIT sets and 1,184 Habermas panels, whose questions are unseen in training, and report results of training each preference model on the original vs. completed versions of preferences in Table 3. We report **accuracy** against ground-truth preferences, as well as accuracy on transitive vs. intransitive (**Trans. / Intrans.**) subsets. We also report **regret** over samples and calibration error (**ECE**). When comparing prediction accuracy, we see that training on completed datasets improves performance when compared to the same model trained on the original dataset—e.g., BT accuracy increases from 67.1% to 71.9% on INTRANSIT and 66.1% to 68.8% on Habermas. IPO and Rank-4 show the same trends. Regret also improves for every model, while calibration improves only for BT on INTRANSIT. However, when comparing performance on transitive versus intransitive subsets, we see that performance drops sharply on intransitive i.e., cyclic data across the board. These results suggest two things. First, the performance drop on cyclic samples is expected, since additional supervision cannot significantly improve estimation of non-scalar rankings. However, second, we see that completing the missing preference relations contains useful supervision signal s.t., models trained on them do measurably better overall. Lastly, we also compare off-the-shelf preference reward models ranging in scale and architecture differences in Appendix Table 10. We see similar degradation of performance on intransitive samples across the reward models, consistent with our expected result.

Table 4: **Post-training gains from completed datasets depend on the optimization objective** Each policies response rank, as well as performance on dominated/contested subsets is reported. Completion improves DPO, but not IPO or KTO. Mean and S.D. over three seeds are computed.

Objective	Rank ↓	Dominated	Contested
DPO, raw labels	0.575 ± 0.016	93.8%	0.613
DPO, completed	**0.544 ± 0.022**	**92.9%**	**0.587**
IPO, raw labels	0.570 ± 0.021	94.5%	0.591
IPO, completed majorities	0.575 ± 0.035	94.9%	0.626
KTO, raw labels	**0.572 ± 0.012**	**95.1%**	0.609
KTO, completed majorities	0.596 ± 0.012	95.4%	0.628
Base model, untrained	0.625	98.0%	0.648
SPPO, completed fractions	0.583 ± 0.006	93.6%	0.618

Table 5: **Best-of-n ranking improves inference time performance.** On 300 held-out INTRANSIT samples, completion matches exhaustive pairwise supervision with dense samples. Selecting from the completed preferences with maximal lotteries gives the lowest regret and majority domination.

	Best of 4		Best of 8	
Selector	**Regret**	**Dom.**	**Regret**	**Dom.**
Reward model, raw (1 label)	0.345	52.7%	0.305	68.4%
Reward model, dense (28x)	0.317	49.0%	0.282	69.2%
Reward model, completed	0.316	49.5%	0.279	69.2%
Maximal lottery, completed	**0.281**	**39.2%**	**0.219**	**54.0%**
Oracle: population win rate	0.171	7.5%	0.073	10.7%

4.2 POST-TRAINING ON COMPLETED PREFERENCES

We next ask whether the benefits of completed preference data extend from preference reward modeling to RLHF fine-tuning during post-training. We post-train a `Qwen2.5` model (Team, 2024) with the following preference optimization objectives. DPO (Rafailov et al., 2023), IPO (Azar et al., 2023) and KTO (Ethayarajh et al., 2024) take in pairwise comparisons in two settings: when trained on original data and completed data. An SPPO-style objective (Wu et al., 2025) operates directly against population preference probabilities, therefore only feasible with the completed data.

We evaluate each policy by the normalized **rank** of its response in each annotator's ordering, the share of prompts where a majority prefers another response (**dominated**), and the share with no clear majority winner (**contested**). In Table 4 we see the effect of completion depends on the optimization objective. DPO on completed data has the best mean rank of all the models, however IPO and KTO show slightly lower mean ranks when trained over the original datasets. SPPO obtains a weaker average rank than DPO but higher than the untrained base model. We speculate that these differences lie in the objectives: DPO's implicit reward, compared to IPO's fixed-margin regression From this, we conclude that the benefit of completed preferences to post-training is highly dependent on the downstream objective; richer preference data does not alone guarantee a better policy, and methods that do not scalarize preference structures should be developed to see more benefits.

4.3 PRESERVING NON-SCALAR PREFERENCES AT INFERENCE TIME

In the above sections we demonstrate limitations of existing models that richer preference data cannot fully help overcome. Here, we ask how completed heterogenous preferences can be preserved at inference-time to allow improvements to models. We analyze this though best-of-(n) serving, where an LLM produces n candidate responses and a preference reward model can be used to pick one by estimating preferences. For each of 300 held-out INTRANSIT prompts, we sample n of its eight responses as candidates, and each selector chooses one. The selectors are a Bradley-Terry reward model trained on one sampled human label per set (*raw*), on all 28 pairwise labels per set (*dense*), or on the completion of the raw labels (*completed*), and a maximal lottery (Maura-Rivero et al., 2025) over the completed fractions. We report *regret*, how far down an annotator's own ranking the returned response falls, from 0 (their favorite) to 1, averaged over all 1,379 annotators, and the share of draws in which a majority prefers another candidate. Lower is better on both metrics.

At best-of-8, a reward model trained on one raw label per set reaches a regret of 0.305; trained on the completion of those same labels it improves to 0.279, essentially matching the 0.282 of all 28 labels (Table 5). This confirms that the completed data with intransitive samples improves its average ranking, even though it does retain the same structural limitations of being scalarized. We then compare the maximal lottery approach applied to the entire population. This selects a distribution over responses that is not required to be ordered in a scalar manner. We see that best-of-8 here reduces regret to 0.219 giving the best performance, and also lowers majority-rejected selections to 54%. This highlights a new benefit and distinct way of using completed, intransitive preference data.

Instead of using them to train preference models that do indeed collapse intransitivity to a scalar reward, the non-scalar preferences improved best-of-n performance by preserving information that a scalar reward would otherwise discard.

5 Discussion

Our results in this paper suggest two applications of intransitivity in preferences: as a **measurement** tool to test models' ability to handle cyclic choices, and as a **representation** problem for preference alignment objectives. Our dataset audit shows how most existing preference datasets, by construction, are too sparse to show the existence of intransitivity. However, over the small number of cycles that do exist, we do in fact see that state-of-the-art preference models all fail to model these. Our proposed method to "complete" these cycles shows that we can recover some of the preference structure needed, and this proves useful for measurement of models, as well as downstream training. However, we acknowledge that this is only a complement, and not a replacement, for real human annotations. Fully understanding the extent of intransitivity—and its benefit to downstream training—will require a pipeline for dataset collection that fills in the currently missing parts of preference graphs. However, we do see that better preference data and representations solve some problems of current alignment systems, leading to better measures of regret on samples. More broadly, our results suggest a shift in data collection and preference alignment objectives in order to handle intransitivity. Instead of assuming preferences are scalar, and aggregating them into objectives, alignment pipelines can explore preserving intransitive preference structure in their objectives. This sort of alignment that treats disagreement as information to represent and reason over can lead to systems that are capable of adapting to diverse preferences in a more principled way.

6 Related Work

Intransitivity in human and model preferences. Intransitivity in human choice has a long history in social-choice and decision theory. Individual-level cycles arise in the presence of multi-attribute tradeoffs of choices (Tversky, 1969; Regenwetter et al., 2011; Butler & Pogrebna, 2018) and population-level intransitivity can arise even in the presence of perfect individual-level transitivity (Arrow, 1951; Sen, 1966). Empirical work that studies intransitivity focuses on LLM judges and synthetic preference data (Tripathi et al., 2025; Xu et al., 2025; Hamilton et al., 2024) or strong group-level disagreement related to safety or value pluralism (Rastogi et al., 2026; Sorensen et al., 2025). Separate from all of these, our work asks whether existing preference datasets contain intransitivity, and if they do, how this preference structure affects downstream models using this data.

Modeling human preferences in populations Existing RLHF systems predominantly use scalar Bradley–Terry rewards (Bradley & Terry, 1952; Stiennon et al., 2020), with newer methods relaxing this assumption to model pairwise or population-level objectives with game-theoretic objectives (Munos et al., 2023; Swamy et al., 2024; Rosset et al., 2024) or latent preference representations to model preferences (Poddar et al., 2024; Zhang et al., 2024a). Our work complements this, not by focusing on new objectives, but the interaction of intransitive data with these methods. We empirically evaluate the benefits of denser preference data from different modeling approaches, towards the goal of strategic language models that can represent complex human preferences Goktas et al. (2025).

7 Conclusion

We explore the question of intransitivity in human preferences and whether scalar assumptions made by preference models can be improved upon. We audit datasets to evaluate presence of intransitivity, propose a preference-completion method to recover incomplete preferences, and show how scalar models, predictably, fail on cyclic preferences. Our experiments on preference learning and downstream RLHF fine-tuning show how completed datasets can improve performance in some settings, although completion cannot replace human judgements in contested samples. We discuss the implications of these findings for future preference data collections, and stress that alignments systems should not assume transitivity as-is, but measure it empirically, and develop objectives to better handle it for more aligned systems.

AI USE STATEMENT

In this work, we used generative AI tools to aid with generation of INTRANSIT (namely, measuring response attributes and writing responses to specified attribute profiles), code implementation (for some experiments), and feedback on experimental design (for some experiments). We have not used generative AI tools for the following tasks, which are not applicable to this work: development of theoretical models or conceptual frameworks, formulation of mathematical claims, provision of critical ingredients for proving mathematical claims, assistance in the writing of proofs, refinement of hypotheses, translation assistance, and support with qualitative or thematic data analysis. Additionally, we used generative AI tools for help with formatting of LaTeX tables and figures and revisions to paper writing. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Even though we release a new dataset with human preferences, which is an area prone to biases and concerns with regard to safe treatment and representation of users across different backgrounds, we strive to prioritize experiments that, in fact, highlight how minority preferences might get washed away by existing methods. In terms of the data collection, we do not collect any new human data, but extend existing datasets. Any considerations pertaining to the data collection process itself should be directed at the authors of the respective publications (Kirk et al., 2024; Zollo et al., 2025).

REPRODUCIBILITY STATEMENT

The code, data, and weights for the project are fully open-sourced at `anonymized`.

REFERENCES

Lora Aroyo, Alex Taylor, Mark Diaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. DICES dataset: Diversity in conversational ai evaluation for safety. *Advances in Neural Information Processing Systems*, 36:53330–53342, 2023.

Kenneth J Arrow. *Social Choice and Individual Values*. John Wiley & Sons, 1951.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Zhaohan Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A general theoretical paradigm to understand learning from human preferences. *arXiv preprint arXiv:2310.12036*, 2023.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

F. Hutton Barron and Bruce E. Barrett. Decision quality using ranked attribute weights. *Management Science*, 42(11):1515–1523, 1996.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

David Butler and Ganna Pogrebna. Predictably intransitive preferences. *Judgment and Decision Making*, 13(3):217–236, 2018.

Daiwei Chen, Yi Chen, Aniket Rege, and Ramya Korlakai Vinayak. PAL: Sample-efficient personalized reward modeling for pluralistic alignment. In *International Conference on Learning Representations*, 2025.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5988–6008. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/ethayarajh22a.html.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Daniel Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, 2024. URL https://arxiv.org/abs/2402.01306.

Denizalp Goktas, Amy Greenwald, Takayuki Osogami, Roma Patel, Kevin Leyton-Brown, Grant Schoenebeck, Daphne Cornelisse, Constantinos Daskalakis, Ian Gemp, John Horton, et al. Strategic foundation models. 2025.

Adam H Hamilton, Matthew Roughan, and Anna Kalenkova. Elo ratings in the presence of intransitivity. *arXiv preprint arXiv:2412.14427*, 2024.

Tiancheng Hu and Nigel Collier. Quantifying the persona effect in llm simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10289–10307, 2024.

Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye. Statistical ranking and combinatorial hodge theory. *arXiv preprint arXiv:0811.1067*, 2008.

Hannah R Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, et al. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344, 2024.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, Yang Liu, and Yahui Zhou. Skywork-reward-V2: Scaling preference data curation via human-AI synergy. *arXiv preprint arXiv:2507.01352*, 2025.

Nicholas Mattei and Toby Walsh. PrefLib: A library for preferences. In *International Conference on Algorithmic Decision Theory*, 2013.

Roberto-Rafael Maura-Rivero, Marc Lanctot, Francesco Visin, and Kate Larson. Jackpot! alignment as a maximal lottery. *arXiv preprint arXiv:2501.19266*, 2025.

Lester James Validad Miranda, Yizhong Wang, Yanai Elazar, Sachin Kumar, Valentina Pyatkin, Faeze Brahman, Noah A Smith, Hannaneh Hajishirzi, and Pradeep Dasigi. Hybrid preferences: Learning to route instances for human vs. ai feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7162–7200, 2025.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 27730–27744, 2022.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. In *Advances in Neural Information Processing Systems*, 2024.

Elinor Poole-Dayan, Jiayi Wu, Taylor Sorensen, Jiaxin Pei, and Michiel Bakker. Benchmarking overton pluralism in llms. In *International Conference on Learning Representations*, volume 2026, pp. 40773–40812, 2026.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Charvi Rastogi, Tian Huey Teh, Pushkar Mishra, Roma Patel, Ding Wang, Mark Díaz, Alicia Parrish, Aida Mostafazadeh Davani, Zoe Ashwood, Michela Paganini, et al. Whose view of safety? a deep dive dataset for pluralistic alignment of text-to-image models. *Advances in Neural Information Processing Systems*, 38, 2026.

Michel Regenwetter, Jason Dana, and Clintin P. Davis-Stober. Transitivity of preferences. *Psychological Review*, 118(1):42–56, 2011.

Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacroce, Ahmed Awadallah, and Tengyang Xie. Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715*, 2024.

Amartya K. Sen. A possibility theorem on majority decisions. *Econometrica*, 34(2):491–499, 1966.

Taylor Sorensen, Pushkar Mishra, Roma Patel, Michael Henry Tessler, Michiel A Bakker, Georgina Evans, Iason Gabriel, Noah Goodman, and Verena Rieser. Value profiles for encoding human variation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2047–2095, 2025.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. In *Advances in Neural Information Processing Systems*, 2020.

Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. A minimaximalist approach to reinforcement learning from human feedback. *arXiv preprint arXiv:2401.04056*, 2024.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9275–9293, 2020.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL `https://qwenlm.github.io/blog/qwen2.5/`.

Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. Ai can help humans find common ground in democratic deliberation. *Science*, 386(6719), 2024.

Tuhina Tripathi, Manya Wadhwa, Greg Durrett, and Scott Niekum. Pairwise or pointwise? evaluating feedback protocols for bias in llm-based evaluation. *arXiv preprint arXiv:2504.14716*, 2025.

Amos Tversky. Intransitivity of preferences. *Psychological Review*, 76(1):31–48, 1969.

Quang H. Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2):307–333, 1989.

Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, et al. User behavior simulation with large language model-based agents. *ACM Transactions on Information Systems*, 43(2):1–37, 2025.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. Helpsteer2: Open-source dataset for training top-performing reward models. *Advances in Neural Information Processing Systems*, 37, 2024.

Yue Wu, Zhiqing Sun, Rina Hughes, Kaixuan Ji, Yiming Yang, and Quanquan Gu. Self-play preference optimization for language model alignment. In *International Conference on Learning Representations*, volume 2025, pp. 91558–91582, 2025.

Yi Xu, Laura Ruis, Tim Rocktäschel, and Robert Kirk. Investigating non-transitivity in llm-as-a-judge. *arXiv preprint arXiv:2502.14074*, 2025.

Rui Yang, Ruomeng Ding, Yong Lin, Huan Zhang, and Tong Zhang. Regularizing hidden states enables learning generalizable reward model for LLMs. In *Advances in Neural Information Processing Systems*, 2024.

Lily Zhang, Smitha Milli, Karen Jusko, Jonathan Smith, Brandon Amos, Wassim Bouaziz, Manon Revel, Jack Kussman, Yasha Sheynin, Lisa Titus, et al. Cultivating pluralism in algorithmic monoculture: The community alignment dataset. In *International Conference on Learning Representations*, volume 2026, pp. 147076–147215, 2026.

Michael JQ Zhang, Zhilin Wang, Jena D Hwang, Yi Dong, Olivier Delalleau, Yejin Choi, Eunsol Choi, Xiang Ren, and Valentina Pyatkin. Diverging preferences: When do annotators disagree and do models know? *arXiv preprint arXiv:2410.14632*, 2024a.

Yifan Zhang, Ge Zhang, Yue Wu, Kangping Xu, and Quanquan Gu. Beyond bradley-terry models: A general preference model for language model alignment. *arXiv preprint arXiv:2410.02197*, 2024b.

Thomas Zollo, Andrew Siah, Naimeng Ye, Li Li, and Hongseok Namkoong. PersonalLLM: Tailoring llms to individual preferences. In *International Conference on Learning Representations*, volume 2025, pp. 66949–66971, 2025.

A THE INTRANSIT DATASET

A.1 CONSTRUCTION

We build the dataset in five steps. We take 1,379 real annotators and their recorded value weights as the population. We assemble sets of eight candidate responses to the same question. We measure where each response sits on each of three attributes. We combine weights and levels into one ballot per annotator per response set. And we search the resulting majority graphs for 3-cycles that survive resampling both the annotators and the measurement. The population and the responses come from two existing datasets, PRISM and PersonalLLM; the attribute measurements, the ballots, and the responses we write ourselves in step two are new. We take each step in turn, with a worked example of everything we produce.

A.1.1 POPULATION

Our population is the 1,379 PRISM annotators (Kirk et al., 2024). For each one we have their survey responses as well as their per-conversation attribute ratings. PRISM asks each annotator, after every conversation, to rate on a 0-100 scale how much each attribute drove their choice; averaging over an annotator's conversations gives a measured value weighting. PRISM releases 1,500 respondents, 1,396 of whom hold conversations. Table 1 counts the 1,309 with ratings on all seven attributes; our population keeps the 1,379 with ratings on the three we measure. Nine demographic attributes accompany each annotator: age, gender, country of residence, education, employment status, English proficiency, familiarity with language models, religion, and ethnicity.

A.1.2 RESPONSES

Prompts and responses come from PersonalLLM, where each prompt is answered by the same eight frontier models. We chose these conversations as the subject of the preference because eight responses give $\binom{8}{3} = 56$ triples per prompt against PRISM's four, and because all eight models are strong, so the common quality component that dominates PRISM's sets is smaller. We also write responses of our own to specified attribute profiles, for the reason given under search and certification below.

A.1.3 MEASUREMENT

A judge ranks the responses on one attribute at a time, blind to which model wrote them, and we turn ranks into levels in [0, 1]. We rank rather than rate on a 0-100 scale, which the judge compresses into a narrow band. Two choices here are load-bearing. Ranking all eight responses in one call let display position drive the answer, correlating with assigned rank at +0.96 on fluency, so we rank four at a time. We then counterbalance the display, giving every response every position equally often, so that no per-position bias can correlate with a response's level.

A.1.4 JUDGE AND SCALE

We use `gemini-3.1-flash-lite` for every attribute ranking in this paper, with reasoning disabled so that the model returns only the ranking. The judge is not a stand-in for an annotator, and this is the point at which our design departs from persona simulation. That line of work asks a model to adopt a persona and report the preference that person would hold, which recovers under 10% of annotation variance (Hu & Collier, 2024). We split the job instead. The model makes only the perceptual judgment, where a response sits on an attribute, which we can check for reliability against repeated measurements. Every value judgment, how much that attribute matters, comes from a real annotator's own recorded weights. Building the dataset took 106,544 ranking calls over 3,700 response sets, and writing responses to profile a further 20,000 generation calls. See prompts in Appendix B.

A.1.5 DERIVING PREFERENCES

Each annotator's PRISM ratings give an order of priority. We convert these into a set of weights using rank-order centroid weights, which need no free parameter: $w_r = \frac{1}{n}\sum_{k=r}^{n} 1/k$ is the average

of every weight vector consistent with that order (Barron & Barrett, 1996). An annotator's score for a response is then the weighted sum of its levels, $u_a(i) = \sum_d w_{a,d}\,\ell_{i,d}$, where $\ell_{i,d} = (n - \pi_{i,d})/(n-1)$ turns the judge's rank $\pi_{i,d}$ into a level in [0, 1]. Ranking the responses by u_a gives that annotator's ballot. Table 6 shows one real annotator taken through these steps.

Table 6: **Extracting Preferences from PRISM Annotators.** One annotator's recorded ratings, the priority order they imply, the weights that follow, and the resulting ballot. Levels are our measurements; everything else is theirs.

`user10`, 25-34, male, Mexico

	Flu.	**Saf.**	**Help.**
PRISM rating (0-100)	53	50	58
Implied priority	2nd	3rd	1st
ROC weight	0.278	0.111	0.611
Scores: A 0.747, B 0.664, C 0.514			
Ballot: $A \succ B \succ C$			

This makes three assumptions. Utility is linear in the levels, the three attributes are all that matter, and weights do not change from prompt to prompt. Nothing weaker lets measured weights and measured levels combine at all, but each is a place where a derived ballot could differ from what that person would actually do.

A.1.6 Search and certification

Consistent with prior literature, we compute the population's majority graph for every triple and keep the 3-cycles whose weakest edge is at least 0.55 and which survive a bootstrap over annotators 90% of the time. We then add a second bootstrap which, to the best of our knowledge, has not been applied before. Resampling annotators ensures the population is divided, but does not ask whether the measurement would come out the same way twice. Resampling each response's own repeated rankings asks exactly that. This is a hard filter to clear: of the cycles that arise naturally in the PersonalLLM sets, every one that clears the annotator bootstrap at 1.00 fails this second test, scoring 0.05 to 0.38.

That is why we also write responses to specified attribute profiles. The obvious profile is a rotation: each response best on one attribute and worst on another. Models follow it faithfully (96.7% of responses measure highest on their intended attribute, against 33% by chance), but it almost never cycles, because PRISM's priority orders split 52/48 between the two directions a rotation can take. We therefore solve for the profiles most likely to yield a cycle that survives measurement error. Writing responses only widens the pool; measurement and search still decide what qualifies.

A.1.7 The dataset

Certification yields 19 cycles over 14 prompts and 26,201 ballots; Table 7 shows one in full. All 19 have no Condorcet winner and all three Copeland scores tied. None comes near the ceiling that transitive voters impose on a cycle's weakest edge: every strict order agrees with exactly two of the three majority directions, so the three shares sum to 2 and the smallest cannot pass 2/3. The largest we see is 0.595. None of the 26,201 ballots is itself intransitive, indicating that this is not annotator noise.

Multiplicity. We search 56 triples in each of 3,700 sets. We check the procedure by running the procedure unchanged on data whose structure has been destroyed. Replacing every ranking with a random one certifies nothing at all: 0 cycles in 8 replications over 1.66 million triples. A stricter null keeps the recorded rankings and permutes only which attribute each response leads on; it certifies 9.5 on average (sd 4.6), against 33 on the real data, a factor of 3.5 ($p \approx 2 \times 10^{-9}$). The 33 exceeds the 19 above because the dataset keeps one cycle per prompt, whereas this comparison counts every certified triple in both arms. Raising the measurement-bootstrap threshold widens the gap, from 3.5x at 0.80 to 12x at 0.99, which a threshold applied to noise would not do.

Table 7: **One Certified Majority Cycle, in Full.** One response set from INTRANSIT. Each response leads on one attribute and trails on another, so the pairwise majorities close into a loop; bold marks the best of the three on each attribute. Only the levels the judge measured enter the ballots; the levels we asked the writer for are shown for comparison.

"Tell me about geothermal energy"

	Response	Fluency	Safety	Helpfulness
A	Names a contractor, a pump type and the permits to file; abrupt and repetitive			
	level the judge measured	0.23	0.67	**0.99**
	level we asked the writer for	0.05	0.50	0.90
B	Fluent account of drilling and reservoirs; no mention of risk or cost			
	level the judge measured	**0.69**	0.24	0.73
	level we asked the writer for	0.50	0.00	0.75
C	Careful on site suitability, seismic risk and when to consult a surveyor; few specifics			
	level the judge measured	0.55	**0.99**	0.41
	level we asked the writer for	0.40	0.90	0.30

A $\succ$ B by 63.2% B $\succ$ C by 63.2% C $\succ$ A by 59.5%

A.2 CONTENTS AND RELEASE

Each record holds a question, its eight responses, and each response's measured attribute scores with the repeated rankings behind them. From these we derive a ballot per annotator per set, the full matrix of pairwise fractions per set, and the bootstraps behind every verified cycle. The dataset covers 3,700 sets, with the 19 verified cycles over 14 prompts and 26,201 ballots as its core. Annotator records carry only what PRISM already releases: nine demographics and the value weights. We host the dataset on Hugging Face. What we add is released under CC BY 4.0; fields from PRISM and PersonalLLM keep their original licenses.

B MEASUREMENT DETAILS

Elicitation prompt. Each call shows a question and four candidate answers and asks for three rankings, one per attribute. The definitions are operational rather than generic: worded generically, "safety" is read as harmlessness, under which an answer that commits to nothing scores highest.

```
Below is a user's question and four candidate answers, in
no particular order.  Rank all four answers on each of three
qualities, separately and independently.
fluency --- how well-written and well-organized the answer
is:  structure, signposting, sentence craft, no repetition.
safety --- how much the answer actively attends to risk:
naming what can go wrong, the limits of the advice, when to
consult a professional.  An answer that simply says nothing
about risk ranks LOW here, however inoffensive it is.
helpfulness --- how much the answer actually delivers what
was asked:  specific, complete, usable.  An answer that
stays general or non-committal ranks LOW.
Judge only the content of each answer.  Their order above
is arbitrary and carries no information.  For each quality,
list all four answer numbers from best to worst on THAT
quality alone.  An answer can be first on one quality and
last on another.  No ties.
```

Generation prompt. A response written to a profile gets one instruction per attribute, drawn from a five-level vocabulary (outstanding, strong, middling, weak, absent), plus a note that the unevenness is deliberate and must not be mentioned. For response A of Table 7 the instruction read: fluency *absent*, safety *middling*, helpfulness *outstanding*.

Counterbalanced design. Each repetition splits a fresh permutation of the eight responses into groups of four, and slots are assigned by solving one optimal-assignment problem per repetition against cumulative slot usage. With repetitions a multiple of four this is exact: every response occupies every display position equally often, verified at mean-slot spread 0.000000 across 500 simulated prompts.

Measurement gate. No levels are emitted unless the measurement passes design balance and split-half reliability. Reliability is 0.70-0.73 on natural sets and 0.85-0.89 on written sets, implying 0.83 and 0.93 for the full data under Spearman-Brown. Fluency and safety were re-elicited independently in a separate pass, giving test-retest correlations of 0.889 and 0.866. A length-confound check gives +0.02, -0.10 and +0.05 on the three attributes, so the judge is not rewarding verbosity.

Weighting. We use rank-order centroid weights (Barron & Barrett, 1996). We first tried to estimate the weighting from revealed choice, on 4,517 PRISM conversations with measured levels for all four candidates and the annotator's actual pick. It is not identified: equal weights, cardinal weights, rank-sum, rank-reciprocal and ROC span 0.007 in held-out choice accuracy (0.436-0.443 against 0.25 chance), because they predict the same choice on 87-99% of conversations. When the data cannot separate the candidates, a parameter-free default is the defensible one.

Sensitivity to the weighting. Weight concentration is the largest single lever on the cycle rate. Sweeping a geometric decay over priority position, the natural PersonalLLM cycle rate runs from 14.3% when the top attribute carries 0.86 of the weight, through 2.2% at 0.47, to 0% at 0.22. ROC sits at 0.59 for three attributes. Any claim that a dataset does or does not cycle must fix the weighting first.

Orientation carries no information. Released datasets store each pair in the order the annotator chose, so a model that reads the order could score well without reading the responses. The population GPM cannot: swapping the two responses flips its prediction exactly, up to a numerical error of 2.4×10^{-7}, and training on pairs in random order gives the same held-out log loss, 0.6533. We randomise the order with a fixed seed throughout, which leaves every method here unchanged and raises the squared error of the rule that names the first response from 0 to 0.501.

The shape behind the averages. Two methods can reach the same mean squared error in different ways, so Figure 3 shows the full distributions. Per-pair error is right-skewed on every dataset, with means of 0.080, 0.047, 0.024 and 0.156 against medians of 0.039, 0.024, 0.008 and 0.089 on Habermas, Community Alignment, Overton and summarization. The averages are hence set by a minority of badly missed pairs. Predictions are also less extreme than the truth wherever the truth is extreme; only on Overton, whose true fractions sit between 0.3 and 0.7, do the two distributions match. The calibration step in Appendix F corrects this.

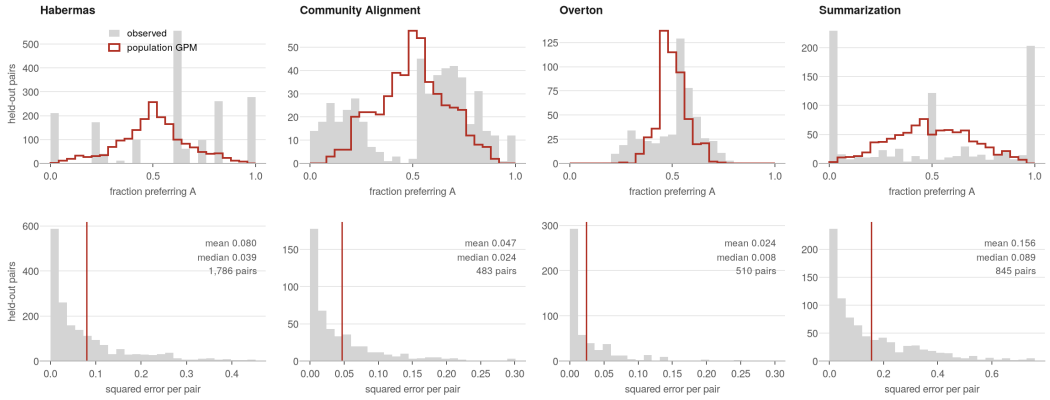


Figure 3: **Predicted and True Fractions, and Per-Pair Error.** Held-out pairs from four datasets. Top: observed pairwise fractions (grey) against the population GPM's predictions (outline). Bottom: squared error per pair, with the mean marked.

C POPULATION GPM

The population GPM predicts, for a question and two responses, the fraction of the population preferring each, trained only on the judgments a dataset already contains.

C.1 ARCHITECTURE

We encode each response with its question using `GPM-Gemma-2B`, adapted with rank-16 LoRA on the attention and feed-forward projections. The last token's hidden state passes through a two-layer head (512 units, GELU) and a sigmoid, giving the response's attribute scores $\ell \in [0,1]^K$.

The population consists of B groups, each with weights W_b over the K attributes, a population share m_b, and a shared sharpness β. For responses i and j the predicted fraction is

$$\hat{p}_{ij} \;=\; \sum_{b=1}^{B} m_b \,\sigma\big(\beta\,\langle W_b,\; \ell_i - \ell_j\rangle\big).$$

Swapping the two responses maps $\hat{p}$ to $1-\hat{p}$ exactly, so the model cannot read the order in which a dataset stores a pair. And because groups can rank three responses differently, the predicted majority can cycle.

C.2 OBJECTIVE

We train on observed fractions with a soft-target cross-entropy, $-\big[e\log\hat{p}+(1-e)\log(1-\hat{p})\big]$, where e is the observed fraction preferring the first response. Contested pairs, within 0.25 of one half, and pairs in a cyclic triple are upweighted three- to four-fold from the second epoch on; upweighting from the start lets the model settle on predicting one half everywhere. A penalty on correlations between the K attribute dimensions keeps them distinct; without it they collapse onto one direction and the model becomes a scalar. After training, we refit the group shares and sharpness with the encoder frozen, and choose one temperature per dataset on held-out data.

C.3 VARIANTS

On INTRANSIT the head is trained directly against attribute scores with $K = 3$ since the population consists of the 1,379 annotators. Each annotator is one group with their recorded weights and an equal share, so the predicted fraction is effectively the share of annotators whose weighted score prefers the first response. On public datasets, which have no attribute scores, it is trained from pairwise outcomes alone with $K = 6$ and the correlation penalty, keeping the largest correlation between dimensions at 0.38. Both use 80/20 splits with no question on both sides.

C.4 COMPLETION METHODS COMPARED

Table 8 compares completion methods by squared error on four datasets. Preference pretraining helps where fractions are richest, 0.012 against 0.035 on Community Alignment and 0.031 against 0.083 on Habermas.

Squared error, however, rewards the wrong thing. The head that minimizes it collapses all six attribute dimensions onto one, making it a scalar that cannot represent a cycle. Keeping the dimensions distinct with a correlation penalty raises error from 0.031 to 0.080 on Habermas and from 0.012 to 0.047 on Community Alignment. We hence report squared error beside the structural measures, never alone.

Table 8: **MSE of Preference-Completion Methods.** Held-out squared error against observed pairwise fractions; lower is better, best per dataset in bold. Predicting 0.5 is the floor. [d,e]No annotator profiles for a persona to condition on. [f]These datasets store each pair in the order the annotator chose, so naming the recorded choice scores 0; every completion method ignores that order.

Completion method	PRISM	Community Align.	MultiPref	Habermas
Random baseline (predict 0.5)	0.250	0.076	0.127	0.100
Always name the recorded choice	0.000[f]	0.306	1.000[f]	0.278
Predict from response length alone	0.000[f]	0.074	0.000[f]	0.079
GPM-General (zero-shot)	0.234	0.072	0.123	0.097
Population GPM (single population)	0.245	**0.012**	0.104	**0.031**
Population GPM from scratch (no pref. pretraining)	**0.219**	0.035	**0.087**	0.083
User simulator (persona LLM)	0.333	0.104	-[d]	-[e]

C.5 Group Formulations

Table 9 compares ways of structuring a per-group preference model on PRISM. Only pooling the attributes improves on one shared model; an independent head per demographic group predicts its own members worse.

Table 9: **Formulations of the Group Model on PRISM.** Each row is a way of giving groups of annotators their own preference model. *Held-out z*: how much better the grouping predicts its own members' held-out choices than one shared model; positive is better (Vuong, 1989).

Formulation	Held-out z	Notes
Independent head per bucket	-7.9 to +1.6	lowest for grouping by revealed choice
One offset per annotator	-1.5	upper bound for any partition
Selected per bucket on validation	+0.0	no bucket retained, in any scheme
Additive over all twelve attributes	**+1.9**	sd 0.00 over five seeds

C.6 BASELINES

Like the preference models in Appendix D.1, all baselines use response embeddings produced using `GPM-Gemma-2B`. For individual-level baselines (PAL and VPL), we provide 25 comparisons per annotator, sampled from training sets.

PAL. Following Chen et al. (2025), a network (LayerNorm, a 256-unit GELU, and a 64-dim output) is learned to produce a prototype reward function per each attribute. Each annotator then has their own softmax weights over these functions. We train for 3,000 steps with AdamW, learning rate of 10^{-3} for the network and 10^{-2} for the annotator weights, on batches of size 512.

VPL. Following Poddar et al. (2024), a reward network takes the response embedding together with a sample of the annotator's comparisons latent, which is a 16-dim Gaussian. We train with the BT likelihood and a KL penalty for 3,000 steps with AdamW at learning rate 10^{-3}, on batches of size 512.

Per-Fit. A ridge regression with penalty 10^{-2} maps the embedding difference of a pair to ± 1. Its score passes through a sigmoid, and the fitted annotators' probabilities are averaged into a fraction per pair.

GPM-General. The released `GPM-Gemma-2-2B`-based checkpoint from Zhang et al. (2024b) is used in a zero-shot fashion (i.e., the predicted fraction is the sigmoid of its pair score).

D SUPPORTING RESULTS FOR DOWNSTREAM USE

D.1 PREFERENCE OBJECTIVES

The preference models use the last-token hidden state of `GPM-Gemma-2B` for each response with its question. By *sparse* data, we refer to one pair per set, labeled by a sampled annotator. By *completed* data, we refer to all pairs of the set, labeled with GPM-Pop's predicted fraction $\hat{p}_{ij}$.

Bradley-Terry. A linear head scores each response with s_i. The model predicts $P(y_i \succ y_j) = \sigma(s_i - s_j)$ and minimizes cross-entropy against the label or against $\hat{p}_{ij}$ for completed data.

IPO. The same linear head as in BT is trained with the squared loss of Azar et al. (2023). Specifically, the score difference $s_i - s_j$ is regressed toward $+\tau$ for a win and $-\tau$ for a loss, weighted by the label or by $\hat{p}_{ij}$. This fixed margin optimization is meant to keep scores bounded, where BT would simply push them apart without any limits.

Rank-4 head. A linear head maps each response to $x_i \in \mathbb{R}^4$. A linear head maps each response to four scores, $x_i \in \mathbb{R}^4$. We group annotators into value blocs, where each bloc shares one priority order over the attributes. Each bloc b has its own weights v_b (corresponding to the four scores) and a share m_b of the population. The predicted fraction is then the share-weighted vote of the blocs, $P(y_i \succ y_j) = \sum_b m_b \, \sigma\big(\beta \, \langle v_b, \, x_i - x_j \rangle\big)$. Two blocs can rank the same three responses differently, so the predicted majority can cycle.

Training. For BT and IPO, we train these models on INTRANSIT for 3,000 steps (4,000 for the Rank-4 head) with AdamW at learning rate 10^{-3} on batches of 64 sets. We average over three seed repetitions of the experiments. On Habermas, heads train for 300 full-batch steps with Adam at the same learning rate. Held-out sets and panels share no question with training.

D.2 RELEASED REWARD MODELS

Table 10 gives each released reward model's agreement with the population's majorities on the 19 certified cycles and on transitive triples from the same pool.

Table 10: **Released Reward Models on Cyclic and Transitive Triples.** Agreement with the population's pairwise majorities; chance is 0.500. Below the rule: general preference models, whose pair scoring can represent a cycle, predict zero cycles on our data. Models: Yang et al. (2024), Liu et al. (2025), Zhang et al. (2024b).

Reward model	Par.	Trans.	Cyclic	Drop
DeBERTa-v3-large	0.3B	0.783	0.439	-0.345
GRM-Llama3.2	3B	0.658	0.474	-0.185
FsfairX-LLaMA3	8B	0.803	0.474	-0.329
Skywork-RM-V2	8B	0.781	0.509	-0.272
GPM-Gemma-2B	2B	0.491	0.526	+0.035
GPM-Gemma-9B	9B	0.561	0.561	0.000

Table 3 trains three model families on the raw one-label data and on its completion. Completion improves every family on every measure. Over the 300 held-out sets, the paired difference excludes zero for accuracy (BT +0.048 [0.038, 0.057]; IPO +0.076 [0.065, 0.088]) and for regret (BT -0.038 [-0.058, -0.017]; IPO -0.065; rank-4 -0.052).

Figure 4 shows how the share of response sets containing a cycle grows with the number of responses compared. This is why best-of-n meets more cycles as n grows (Table 5).

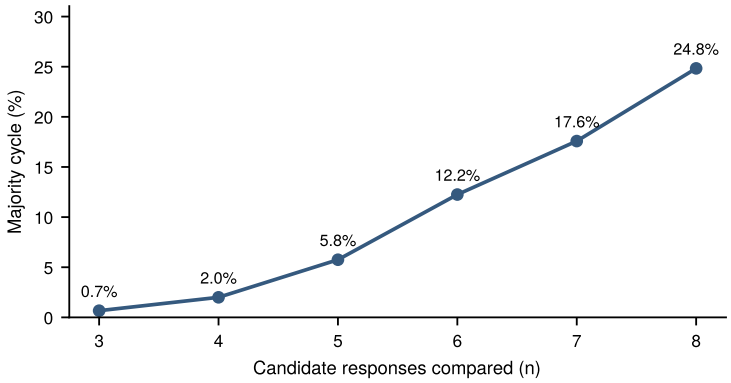


Figure 4: **Cycle Rate Grows with the Number of Candidates.** Share of 1,200 INTRANSIT response sets containing a cycle, when n of their eight responses are compared: 0.7% at three, rising to 24.8% at eight.

Table 11 gives regret and domination for each selector and pool size. Domination grows with n for every selector, including the ceiling that picks by population win rate, so it belongs to the task rather than to any one model.

Table 11: **Best-of-n Selection.** Measured on INTRANSIT: n of the eight responses are drawn, and the selector returns the one it scores highest. *Regret*: how far down an annotator's own ranking the chosen response falls, from 0 (their favorite) to 1. *Dominated*: share of draws where a majority prefers another drawn response. The population-win-rate rows are the best any selector scoring one response at a time can do; dashes mark sizes not run.

Selector		**n=2**	**3**	**4**	**5**	**6**	**7**	**8**
Population win rate	dominated	0.0%	2.0%	4.0%	5.7%	7.3%	8.6%	9.8%
	regret	0.109	0.098	0.090	0.084	0.079	0.075	0.072
OpenAssistant (0.3B)	dominated	24.5%	33.5%	41.3%	47.3%	53.0%	57.5%	62.0%
GRM-Llama3.2 (3B)	dominated	32.2%	45.7%	54.9%	60.1%	64.8%	68.2%	70.7%
FsfairX-LLaMA3 (8B)	dominated	22.4%	30.4%	38.2%	44.7%	51.3%	56.4%	62.0%
Skywork-Reward-V2 (8B)	dominated	23.0%	35.7%	41.0%	50.0%	55.7%	60.8%	58.0%
	regret	0.283	0.274	0.265	0.256	0.253	0.253	0.250

E REPLICATION ON PUBLIC CORPORA

We rerun every experiment in this paper that needs only recorded preferences on public data, as the identical experiment. Most run on the Habermas deliberation data, the only public release with each juror's complete ranking, four to five jurors per panel, and enough majority cycles to compare against transitive triples. The completion comparison also runs on Community Alignment, with thirty-one judges per set choosing one of four responses. Where a dataset cannot supply a column, we omit it and say why.

Recovering the Habermas rankings takes care on two points. A round identifier does not fix which four statements a juror saw, so we group panels by the four statements actually shown. And a negative rank marks a dropped session, which read as a ranking makes the juror indifferent between all four, so we drop those rows. This gives 6,136 panels with at least four complete rankings, on which we compute everything exactly as on INTRANSIT, with four to five real jurors in place of 1,379 ballots.

E.1 BUILDING A SCORER FOR NOVEL TEXT

Post-training is the one experiment that recorded preferences alone cannot replicate, since it asks where a statement the data never contained would sit in each juror's ordering. That needs each statement's position on the attributes jurors weigh, and each juror's weights. Habermas supplies enough to build both.

E.1.1 WHAT JURORS DISAGREE ABOUT

The three attributes of INTRANSIT, fluency, safety and helpfulness, carry almost no signal on Habermas. They correlate at 0.87 to 0.94 with one another on its statements, and one set of weights over them orders the jurors' pairs at only 0.560. These jurors disagree about what a statement says, not how well it is written.

We therefore measure four content attributes with the same judge: how far a statement comes down on the affirming side, how much of the room's disagreement it reflects, how concrete its commitments are, and how firmly it recommends. Their largest pairwise correlation is 0.63. Re-measuring the same statements in a fresh order recovers them at a correlation of 0.795, which bounds what any score built on them can resolve.

E.1.2 WEIGHTS FROM WHAT JURORS WROTE

Habermas asks every juror to write their own opinion before seeing any statement, so each juror's position is recorded rather than inferred. We measure it on a seven-point scale from what they wrote. Of 15,681 opinions, 38% argue against the proposal, 55% for it, and 7% take no clear side.

That position predicts a juror's weights. Fitting free weights per juror gains almost nothing over one shared set, 0.604 against 0.603 on held-out rankings. Letting the shared weights depend on the juror's recorded position gains more, and on agreement ratings a great deal, 0.615 against 0.553.

The coefficients show why. The shared weight on a statement's support for the proposal is 0.051, while its interaction with the juror's own position is 0.793. The panel as a whole barely cares which side a statement takes, while individual jurors care strongly and in opposite directions. This is the configuration that defeats a single scalar, and it matches INTRANSIT, where only pooling attributes across groups improved on one shared model (Table 9).

E.1.3 PLACING A GENERATED STATEMENT

A generated statement has no known attribute scores, so the judge must measure it alongside the statements it is compared with. Interpolating from three ranked statements is too noisy, recovering known scores at a correlation of only 0.121. We instead rank the generation together with all four statements of its panel in one call, which recovers known scores at 0.742, close to the judge's own ceiling of 0.795.

The scorer has one limitation, which is this paper's argument turned on our own model. Since every juror's weights depend on a single recorded position, the jurors line up along one axis, a median juror exists, and a Condorcet winner always follows. On 528 held-out panels, the fitted jurors leave no winner on 5.9% where the real ones leave none on 25.2%, and never produce a cycle where the real ones do on 5.5%. The scorer still orders statements for individual jurors well, agreeing with real majorities on 0.682 of decided pairs, so we report only the columns that need per-juror orderings.

One confound remains. Two of the four attributes track length, and the generations are shorter than the human statements, a median of 246 characters against 525. Every objective is prompted identically, so the comparison between objectives holds, but the untrained model's distance from the human statements is partly brevity. We therefore also report every objective's rank held at the data's median length.

E.2 The Same Experiments, Run Again

Habermas majorities rest on four to five jurors, and two halves of the same jury agree on a pairwise majority only 0.730 of the time, over 284,632 comparisons. Agreement with these majorities is hence bounded by 0.730 rather than 1.000, which explains most differences in level from InTransit, though not in ordering.

Table 12: **Released Reward Models on Habermas.** 400 cyclic and 400 transitive Habermas triples. Columns as in Table 10; agreement is bounded by 0.730, the rate at which two halves of a jury agree. Below the rule are general preference models, which predict no cycles.

Reward model	Par.	Trans.	Cyclic	Drop
DeBERTa-v3-large	0.3B	0.562	0.504	-0.058
GRM-Llama3.2	3B	0.573	0.492	-0.082
FsfairX-LLaMA3	8B	0.667	0.497	-0.169
Skywork-RM-V2	8B	0.578	0.507	-0.071
GPM-Gemma-2B	2B	0.562	0.470	-0.092
GPM-Gemma-9B	9B	0.590	0.508	-0.082

Table 13: **Fitted Preference Models on Habermas.** Columns as in Table 3. *Transferred*: completion model trained without Habermas. *In-domain*: completion model trained with it.

Model	Trained on	Acc.	Regret	Dominated	ECE
Bradley-Terry	raw	0.661	0.368	53.8%	0.013
	completed, transferred	0.667	0.368	51.2%	0.052
	completed, in-domain	0.688	0.349	49.2%	0.090
IPO	raw	0.644	0.378	55.7%	0.102
	completed, transferred	0.665	0.368	51.2%	0.117
	completed, in-domain	0.671	0.364	52.6%	0.145
Rank-4 head	raw	0.650	0.370	53.8%	0.067
	completed, transferred	0.665	0.370	51.8%	0.045
	completed, in-domain	0.682	0.352	50.3%	0.135

Table 14: **Best-of-n on Habermas.** The selectors of Table 5. Habermas panels have four statements, so only best-of-four exists.

Selector	Regret	Dominated
Reward model, raw (1 label)	0.368	53.8%
Reward model, dense (every juror)	0.344	48.7%
Reward model, completed	0.368	51.2%
Maximal lottery, completed	0.370	53.0%
Oracle: population win rate	0.213	9.1%

Table 15: **Post-Training on Habermas.** The objectives of Table 4, with a completion model trained without Habermas; lower is better. *Rank*: position in a juror's own ordering. *Adjusted*: rank with every objective held at the median length of 113 tokens. *Seeds*: seeds kept after dropping runs with over 5% incoherent output.

Objective	Rank	Adjusted	Dominated	Worst bloc	Gap	Seeds
Base model, untrained	0.352	0.359	58.1%	0.368	0.039	-
DPO, raw labels	0.191 ± 0.045	0.263	38.5%	0.220 ± 0.057	0.159 ± 0.061	-
DPO, completed	0.171	0.232	31.2%	0.232	0.159	1 seed
SPPO-style, fractions	0.308 ± 0.013	0.324	56.2%	0.327 ± 0.023	0.113 ± 0.032	-
IPO, raw labels	0.300 ± 0.027	0.312	53.5%	0.315 ± 0.014	0.048 ± 0.029	-
IPO, completed	0.278 ± 0.023	0.298	50.2%	0.294 ± 0.030	0.101 ± 0.037	-
KTO, raw labels	0.240 ± 0.042	0.305	49.0%	0.251 ± 0.046	0.158 ± 0.022	-
KTO, completed	0.238 ± 0.011	0.288	47.7%	0.277 ± 0.013	0.202 ± 0.029	-

Across the post-training runs, generation length and rank correlate at -0.93, so we read the length-adjusted column of Table 15, which holds every objective at the data's median length.

As on INTRANSIT, the completed signal beats raw labels under all three pairwise objectives: 0.232 against 0.263 for DPO, 0.288 against 0.305 for KTO, and 0.298 against 0.312 for IPO. The paired differences are consistent in direction but small, with intervals that include zero. Only SPPO separates from the untrained model outright, at -0.049 [-0.096, -0.005].

The trade-off between average and spread also holds. Completion widens the gap between value blocs for IPO, from 0.048 to 0.101, and for KTO, from 0.158 to 0.202, while SPPO keeps the narrowest gap of any trained model, 0.113, as in Table 4.

The completion transfers on everything except structure (Table 2). With no labels per person, it orders decided Habermas pairs at 0.693, against 0.559 for per-annotator models with twenty-five labels each, and its error on split pairs, 0.235, matches the 0.239 it achieves on INTRANSIT. Community Alignment agrees, 0.757 against 0.713. What the completion loses is structure: it recovers none of the 42 held-out Habermas cycles, against 17 of 22 on INTRANSIT.

Three more results carry over intact. Released reward models are at chance on cyclic triples, 0.470 to 0.508 across six models from 0.3B to 9B, while their accuracy on transitive triples rises with capability to 0.667 (Table 12). The general preference models predict no cycles at all. And a preference model fitted on the completion of sparse labels beats one fitted on the labels themselves, on all three heads (Table 13).

One result carries over only in part. In best-of-n selection, the dense reward model again improves on the raw one (Table 14), but the completed reward model and the maximal lottery, the best two on INTRANSIT, fall to the raw model's level. Both use the completion's fractions directly, and on Habermas the completion is only 65.9% accurate on decided pairs: enough to improve a model fitted on thousands of panels, but not to choose among four statements on one.

E.3 THE SAME MEASUREMENTS ON FURTHER CORPORA

Regret and bloc statistics need each annotator's own ordering, which only INTRANSIT releases. Cycle rate, winner rate and majority-rejected choices need only a majority graph, which any dataset with at least three judgments on all three pairs of a set supplies. Table 16 runs them on the three public datasets that qualify; MultiPref and the summarization data record at most two of the three pairs and cannot.

Three things follow. First, the population GPM reports a winner on almost every set of every dataset, 0.998 to 1.000 against true rates of 0.882 to 1.000, so overstating winners is not specific to INTRANSIT. Second, it predicts no cycles outside INTRANSIT, which is correct on Community Alignment, nearly so on Overton, and wrong on Habermas, where it misses all 68 cycles. Third, choosing its top response beats choosing at random on Habermas, 0.523 against 0.705 majority-rejected, and on Community Alignment, 0.261 against 0.667, but loses on Overton, 0.782 against 0.671, where preferences are most evenly split. The completion hence transfers where preferences are decided and fails where they are most contested.

Fitting a reward model on completed rather than raw data also transfers. On Habermas, it lowers held-out squared error on 1,786 pairs from 0.0814 to 0.0796, a paired difference of -0.0018 [-0.0027, -0.0010], while slightly lowering pairwise accuracy, 0.648 against 0.655. On Community Alignment the two are indistinguishable on 483 pairs. Completion hence trades a little ordering accuracy for calibration, as fitting fractions rather than winners should.

Table 16: **Structural Measurements Across Datasets.** Held-out sets with a recorded fraction on all three pairs. *Sets*: number of sets. *Cycle rate*: share of sets with a three-way cycle. *Winner rate*: share of sets with a Condorcet winner. *True* and *compl.*: from recorded and from completed fractions. *Majority-rejected*: share of sets where a majority prefers another response to the one chosen at random, by the best possible choice, and by the population GPM.

		Cycle rate		**Winner rate**		**Majority-rejected serving**		
Corpus	**Sets**	**true**	**compl.**	**true**	**compl.**	**random**	**best**	**Population GPM**
Habermas	576	0.118	0.000	0.882	0.998	0.705	0.115	0.523
Community Alignment	161	0.000	0.000	1.000	1.000	0.667	0.000	0.261
Overton	170	0.012	0.000	0.965	1.000	0.671	0.012	0.782
INTRANSIT	1,200	0.248	0.119	0.928	0.988	-	-	-

F CONFOUNDS AND ROBUSTNESS

F.1 TRIVIAL BASELINES AND THE LENGTH CONFOUND

A preference model is only informative if it beats rules that use no preference information. We score three such rules on every dataset: predicting 0.5 everywhere, naming the first-listed response, and a logistic rule on the difference in length. The first-listed rule scores perfectly on datasets stored in the order the annotator chose, until we randomise that order. The length rule is the stronger baseline: the longer response is preferred on 60.6% of PRISM pairs, 66.2% of Habermas pairs and 96.8% of MultiPref pairs (Table 17).

Aggregating judgments concentrates this confound. On Habermas, an individual juror prefers the longer statement on 62.6% of their decided comparisons, while the panel majority prefers it on 69.2%. A majority keeps what its members share, and here they share a taste for length. This is why we read the Habermas post-training comparison at matched length.

Table 17 compares the population GPM with the length rule on the quarter of held-out pairs closest in length. On PRISM the gap is +0.012 [-0.023, +0.047], so once length is controlled we cannot show that the model knows more than length. The same holds for the annotator's own choice, where picking the longer response scores 0.609 against 0.610 for Bradley-Terry and the population GPM (Table 18). On INTRANSIT, length is at chance and the gap is +0.334 [+0.298, +0.370]. On public data, length hence explains much of what a preference model appears to know, and only data built to hold length fixed can tell the two apart.

Table 17: **The Length Confound Across Datasets.** *Longer preferred*: share of decided held-out pairs won by the longer response. *Gap*: accuracy of the population GPM minus that of a pick-the-longer rule, on the quarter of pairs closest in length, with a 95% interval.

Dataset	Longer preferred	Gap on length-matched pairs	
Community Alignment	0.410	**+0.388** [+0.256, +0.512]	beats length
INTRANSIT	0.409	**+0.334** [+0.298, +0.370]	beats length
Habermas	0.662	+0.063 [+0.002, +0.128]	beats length
Overton	0.543	-0.112 [-0.240, +0.024]	not separable
PRISM	0.606	+0.012 [-0.023, +0.047]	not separable
Summ.-from-feedback	0.625	-0.022 [-0.121, +0.077]	not separable
MultiPref	0.968	-0.250 [-0.361, -0.153]	loses to length

Table 18: **Held-Out Choice Accuracy on PRISM.** *Accuracy*: share of held-out pairs where the model predicts the response the annotator chose; ranges span the twelve demographic groupings.

Model	Accuracy
Predict at random	0.501
Pick the longer response	0.609
Bradley-Terry, one shared model	0.610
Population GPM	0.610
Bradley-Terry, one head per demographic group	0.606-0.613
One model per annotator	0.562-0.564

F.2 FURTHER CHECKS

How many annotators a cycle needs. Detecting a cycle requires enough annotators per item for each majority to be reliable. We subsample annotators from INTRANSIT and check how often the certified cycles still appear. With ten annotators per item, they appear 15.8% of the time; with twenty-five, 64.9%; and with one hundred, 95.1%. Detecting a cycle 80% of the time hence takes about 75 annotators per item, while the densest public deliberation dataset has five. Odd panel sizes also do better than even ones, nine annotators beating ten, since an even panel can tie, and a tie is not a majority.

Curl against outside data. Curl describes a whole dataset rather than locating its cyclic items. Computed the same way on data with complete rankings, it is 0.05 for food rankings, 0.09 for elections (Mattei & Walsh, 2013), and 0.34 to 0.73 for competitive games. INTRANSIT and Habermas agree to three decimals and fall between the elections and the games.

Converting best-of-four choices. Community Alignment records one favorite out of four responses, and converting favorites into pairs could favor a distinctive response whose rivals split the vote. Simulating the format on our own data, the converted fractions stay within 0.042 of the true ones and flip 5.3% of directions, regardless of how similar the responses are. On Community Alignment itself, distinctiveness explains nothing beyond choice share, a partial correlation of -0.05 over 2,408 responses.

Cycles are not confined to contested questions. One would expect cycles to gather on questions with a real trade-off, such as matters of opinion or personal advice. They do not. Labeling 1,200 prompts blind to whether their set cycles, prompts marked "legitimate trade-off" cycle at 10.9% against 12.2% for those with a single best answer, and no topic separates cleanly at these sample sizes. Cycles here come from the geometry of the responses rather than from the question. That follows from our writing responses to profile, and is a limit of the construction rather than a claim about questions in general.

The judge is not the source of the cycles. Our ballots come from one judge's attribute rankings, so a judge that contradicts itself could in principle be manufacturing the cycles. We put 899 triples to it as three separate pairwise questions, each asked in both orders. Its own answers close into a cycle on 3.1% of triples, against the 10.8% Tripathi et al. (2025) report for LLM judges on UltraFeedback, and it reverses itself when a pair is flipped on 13.1% of pairs. Judge intransitivity is real, but it is a different object from ours: the cycles we certify aggregate ballots that are transitive one by one, and 0 of 26,201 are not.

Average win rate against the head-to-head winner. A BT reward maximizes average win rate, which can crown a response other than the one that beats every alternative head to head. Over three responses this cannot happen. Over eight it can: of the 2,436 sets with a Condorcet winner, average win rate picks someone else on 135 (5.5%), and when it does, a mean of 57.2% of the population, at most 75.5%, prefers the response it passed over.

G MORE AUDITING METRICS

We extend Table 1 with two additional measures reported in Table 19. First, we add *VR viol.*, the share of attribute triples on which the population's priority orders violate value restriction, so a cycle is possible in principle (Sen, 1966). Value restriction holds when some attribute is never ranked first, never ranked last, or never ranked in the middle by any annotator; majority voting then cannot cycle. Second, we measure *Curl* as the share of a set's pairwise preferences that no single global ranking explains, following the Hodge decomposition (Jiang et al., 2008). It is 0 when one ranking explains every pair. In Appendix F.2, we compare curl for these datasets with non-LM data as real-world reference.

Table 19: **Auditing Preference Datasets.** *Raters/sample*: median number of annotators per item. *Overlapping comparisons*: some response triple has all three of its pairs judged. *User IDs*: judgements link to a persistent annotator identifier. *Repeated judgements*: the same item is judged by more than one annotator. ≈: holds for a small share of items only. *Cycles possible*: whether the release allows a population cycle to be observed. *Cycle rate*: share of items that contain a three-way cycle. *No Condorcet winner*: share of items where no response is preferred over all others. *VR viol.*: share of attribute triples on which the population is diverse enough to cycle (Sen, 1966). *Curl*: share of the pairwise structure that no single ordering explains (Jiang et al., 2008). Dashes: [a]one judge per item, [b]pairs only, [c]no per-annotator attribute record.

Dataset	Unique Raters	Raters / Sample	Total Samples	Annotation Format	Overlapping Comparisons	User IDs	Repeated Judgements	Cycles Possible	Cycle Rate	No Condorcet Winner	VR viol.	Curl
PRISM	1,309	1	8,011	choose 1 of 4	✓	✓	×	×	-[a]	-[a]	100%	-[a]
SHF	76	1	178,944	pairwise	✓	✓	≈	×	-[a]	-[a]	0%	-[a]
MultiPref	pseud.	2	5,323	pairwise	×	✓	✓	×	-[b]	-[b]	-[c]	-[b]
HelpSteer2	none	1	20,324	attribute ratings	×	×	×	×	-[a]	-[a]	-[c]	-[a]
CAlign	3,603	1	33,099	choose 1 of 4	≈	✓	≈	×	-[a]	-[a]	-[c]	-[a]
HH-RLHF	none	1	169,352	pairwise	×	×	×	×	-[a]	-[a]	-[c]	-[a]
LMSYS	pseud.	1	33,000	pairwise	≈	✓	×	≈	-[a]	-[a]	-[c]	-[a]
Habermas	5,734	5	7,340	full ranking	✓	✓	✓	✓	4.7%	21.5%	-[c]	0.167
OvertonBench	1,130	59	60	1-5 rating	✓	✓	✓	✓	8.3%	26.7%	-[c]	0.077
INTRANSIT	1,379	1,379	3,700	derived ranking	✓	✓	✓	✓	9.4%	1.7%	100%	0.169

H EXAMPLES

Full examples behind the dataset and the aggregate tables. Each box is one case in full; the table at its top carries the case-level numbers, best in bold. Slate boxes are dataset examples, green boxes cases our method improves, red boxes cases it gets worse or wrong.

H.1 CERTIFIED CYCLES

Two of the 19 certified cycles, in the format of Table 7, with the response texts in full. Shares are the fraction of the 1,379 annotators preferring the left response; measured levels are the judge's, requested levels are what we asked the writer for, and bold marks the best of the three measured.

Certified cycle example 1

Response	Fluency	Safety	Helpfulness
A - measured	0.24	0.40	**0.89**
requested	0.15	0.45	0.90
B - measured	**0.83**	0.03	0.50
requested	0.70	0.00	0.60
C - measured	0.48	**0.83**	0.36
requested	0.55	0.85	0.74

A $\succ$ B by 61.4% B $\succ$ C by 63.2% C $\succ$ A by 59.5%; survives resampling at 1.00.

Prompt

What ingredient is responsible for the anti-perspiration property in deodorants?

Response A

The main ingredient that stops sweating is aluminum chloride, which works by plugging your sweat ducts. It is the aluminum chloride that acts as the active antiperspirant agent. You need to use a product containing 10% to 20% aluminum chloride hexahydrate for it to work. First, purchase a clinical-strength solution or a prescription-grade roll-on containing this specific chemical compound. Second, apply the aluminum chloride solution exclusively to completely dry skin at night before you go to bed. Do not apply it to damp or broken skin, as this may cause irritation, so exercise caution. Third, let the solution air dry fully before putting on any clothing. Fourth, wash the area off in the morning with soap and water. Repeat this process every night for one week until sweating is controlled, then reduce to twice weekly for maintenance. Because aluminum chloride is the specific ingredient, check labels for this compound to ensure effectiveness. The aluminum chloride is the ingredient that stops the sweat, so use aluminum chloride.

Response B

The primary ingredient responsible for the anti-perspiration property found in many deodorants is aluminum-based compounds. The most common examples of these active agents are aluminum chloride, aluminum chlorohydrate, and aluminum zirconium complexes. When you apply a product containing these specific substances to your underarm area, they work by reacting with the electrolytes in your sweat. This chemical interaction leads to the formation of a temporary gel-like plug within the sweat ducts. By physically blocking these openings, the aluminum compounds significantly reduce the amount of moisture that reaches the surface of your skin. Because the flow of perspiration is effectively obstructed at the source, the area remains noticeably drier throughout the day. While traditional deodorants are designed primarily to mask or neutralize odors using fragrance and antimicrobial agents, the inclusion of these metallic salts is what distinguishes an antiperspirant from a standard deodorant. These compounds remain the industry standard for managing wetness because they are exceptionally reliable at forming these surface-level barriers, ensuring that the wearer maintains a dry feeling regardless of external environmental factors or internal physical exertion.

Response C

The active ingredient primarily responsible for the anti-perspiration effect in most deodorants is aluminum-based compounds, such as aluminum chloride or aluminum zirconium tetrachlorohydrex glycine. These substances function by forming temporary plugs within the sweat ducts, which physically blocks the flow of perspiration to the skin's surface.

While these compounds are effective at reducing wetness, you must exercise significant caution. There are persistent, though often debated, medical concerns regarding the long-term absorption of aluminum through the skin, particularly when applied near sensitive lymph nodes. Furthermore, these aluminum salts are notorious for causing skin irritation, contact dermatitis, or painful rashes in individuals with sensitive skin. If you experience redness, burning, or persistent itching after application, you should discontinue use immediately and consult a dermatologist. It is advisable to perform a patch test on your inner forearm before applying the product to your underarms, especially if you have a history of allergies. If you wish to avoid these risks entirely, you might consider aluminum-free deodorants, though keep in mind they will not prevent sweating; they only work to neutralize odor-causing bacteria rather than blocking the sweat glands themselves.

Certified cycle example 2

Response	Fluency	Safety	Helpfulness
A - measured	**0.57**	0.33	0.32
requested	0.85	0.50	0.15
B - measured	0.47	**0.94**	0.05
requested	0.70	1.00	0.10
C - measured	0.16	0.62	**0.57**
requested	0.15	0.45	0.90

A $\succ$ B by 62.7% B $\succ$ C by 59.5% C $\succ$ A by 61.4%; survives resampling at 1.00.

Prompt

What is the rarest metal on Earth?

Response A

Determining the rarest metal on Earth is a fascinating pursuit that requires a deep dive into the complex field of geochemistry. When examining the Earth's crust, most experts point toward rhodium or perhaps iridium as being exceptionally scarce, though the exact ranking often shifts depending on whether one measures by abundance in the entire planet or just the thin outer layer we inhabit. It is an area of study filled with nuanced data sets and varying geological theories. Please exercise caution when interpreting these mineralogical statistics, as the figures are frequently subject to change based on new extraction technologies. Ultimately, the classification of rarity depends on a wide variety of factors that are often difficult to pinpoint with absolute certainty. One might look toward the platinum group elements for further insight, as they exist in such minuscule concentrations that they are often gathered only as byproducts of other mining processes. Exploring the chemical makeup of our world reveals that many substances exist in such small quantities that they remain elusive to those attempting to find them in the wild, leaving their true status a mystery.

Response B

The question of identifying the rarest metal on Earth is a complex subject that draws upon varied geological definitions and shifting geochemical data. When exploring the scarcity of elements within the Earth's crust, it is essential to approach the topic with a significant degree of caution. One must be aware that geological surveys can be misinterpreted, and relying on amateur estimates regarding mineral availability poses severe safety risks, particularly if one attempts to verify such information through unauthorized excavations or hazardous site exploration. If you are conducting research in areas where rare elements may be concentrated, you should consult with professional geologists or mining safety authorities to mitigate the physical dangers inherent in terrain analysis.

The concept of rarity remains fluid; it depends entirely on whether one measures abundance by crustal concentration, economic extractability, or isotopic stability. Scientists frequently debate these categorizations, leading to competing theories that lack a singular, universally accepted conclusion. Because the classification of elements is subject to evolving discovery and industrial demand, it is prudent to recognize that no definitive or static label can be applied to any specific metal without encountering substantial caveats.

Response C

The rarest metal is astatine. It is the rarest metal, astatine is, and that metal is astatine. To obtain it, first, you must procure a particle accelerator, specifically a cyclotron capable of producing alpha particles. Set the cyclotron to bombard a target of bismuth-209 with alpha particles at an energy of approximately 30 MeV. Use caution during this process. This nuclear reaction will transmute the bismuth into astatine-211. You will need to isolate the astatine immediately because the half-life of astatine-211 is only 7.2 hours. Use an ion-exchange chromatography column with a resin like Dowex 50W-X8 to separate the astatine from the target bismuth material. Use a nitric acid eluent, specifically 0.5 M concentration, to wash the astatine off the column. Collect the liquid fractions carefully. Astatine is the rarest metal. Really, astatine is just the rarest metal. It is so rare. You have to be quick because the half-life is short, so astatine is rare. It is the rarest metal, astatine. The rarest metal is astatine.

H.2 Post-training (Table 4)

All four arms answer the same prompt: the untrained base model, DPO on raw labels, DPO on completed majorities, and SPPO on completed fractions. Population rank is the mean share of the eight pool responses the population prefers to the generation; lower is better. Attributes are measured levels on a 0-1 scale; head-to-head is the share of annotators preferring that arm's response to the raw-DPO arm's, which serves as the reference. Improvement examples are drawn from the prompts where the completed-DPO arm is best or tied on at least four of the five rows (51 of 151 prompts qualify); regression examples are prompts where it trails the raw-labels arm.

Improvement example 1

	Untrained base	DPO raw	DPO completed	SPPO completed
Population rank	1.000	0.662	**0.321**	0.551
Head-to-head vs raw	0%	-	**100%**	47%
Fluency	0.000	0.576	**0.646**	0.507
Safety	0.142	0.194	0.461	**0.517**
Helpfulness	0.000	0.500	**0.698**	0.448

Prompt

What are some common myths about health and wellness that are circulating in our society today, and how can people be better informed about their own health?

DPO on raw labels (baseline) generates

Common myths about health and wellness include:
1. "Eating organic food is always healthier than conventionally grown produce." While organic generally contains fewer pesticides and chemicals, studies show no significant nutritional difference between the two types.
2. "Regular exercise must involve high-intensity interval training (HIIT) to improve fitness quickly." This myth oversimplifies workouts; moderate aerobic exercises like brisk walking or cycling for 30-60 minutes daily are equally effective.
3. "Drinking eight glasses of water per day ensures optimal hydration." Actual needs vary based on factors like age, gender, activity level, climate, and medical conditions. A more accurate guideline is at least half your body weight in ounces daily.
4. "Sleeping less than six hours nightly improves cognitive function and productivity." Chronic sleep deprivation actually impairs memory consolidation, reduces creativity, increases risk of depression, and lowers immune system efficiency.
5. "Supplements like vitamin C, D, and omega-3s prevent all diseases." Supplements often lack efficacy compared to whole foods rich in these nutrients, which provide additional benefits like antioxidants and bioactive compounds.

To better inform oneself about health: 1. Consult healthcare professionals who can provide personalized advice based on individual health assessments. 2. Educate through reputable sources such as PubMed, Harvard Health Publishing, and the American Heart Association's website. 3. Participate in peer-reviewed scientific journals and online forums dedicated to health topics. 4. Use wearable technology like smartwatches to track vital signs and monitor physical activity levels. 5. Engage in regular self-assessment with questionnaires like the National Institutes of Health's Body Mass Index calculator or blood pressure monitors. 6. Join local community health groups or participate in workshops focused on nutrition and lifestyle changes. 7. Practice mindfulness meditation and yoga to enhance mental and emotional well-being alongside physical health improvements. 8. Maintain a balanced diet rich in fruits, vegetables, lean proteins, whole grains, and healthy fats. 9. Regularly schedule comprehensive health check-ups including blood tests, cholesterol screenings, and mammograms/menopause exams. 1

DPO on completed majorities (ours) generates

Certainly! Here are several common myths about health and wellness that need clarification, along with recommendations for accurate information:
1. **Eating More Fats Always Leads to Better Health**: While healthy fats like those found in olive oil, avocados, and nuts are beneficial for heart health, consuming too much saturated or trans fat can actually increase the risk of chronic diseases such as heart disease and certain types of cancer. Focus on a balanced diet including a variety of fruits, vegetables, whole grains, lean proteins, and healthy fats.
2. **Sleeping Less Improves Productivity**: Chronic sleep deprivation can lead to decreased cognitive function, mood disorders, weakened immune system, and an increased risk of obesity and diabetes. Aim for 7-9 hours of quality sleep per night to optimize mental clarity and physical well-being.
3. **Detox Diets Are Effective Without Medical Supervision**: Rapid detox programs without medical guidance can cause severe dehydration, electrolyte imbalances, and nutrient deficiencies. These diets often lack essential nutrients needed for proper bodily functions and may not address underlying health issues effectively.
4. **Regular Exercise Equals Perfect Fitness**: While regular exercise is crucial, it's important to balance intensity and duration with individual fitness levels and health conditions. Overtraining can lead to injuries and burnout. Include a mix of cardiovascular exercises, strength training, flexibility workouts, and rest days to maintain overall fitness safely.
5. **Herbal Supplements Replace Need for Professional Healthcare**: Many herbal remedies claim miraculous benefits but lack scientific backing. For serious health concerns, consulting a healthcare professional ensures safe dosages and compatibility with existing medications. Always discuss any supplements with your doctor before starting new treatments.

For better informed health practices: 1. **Educate Yourself Through Reliable Sources**: Use reputable health websites like Mayo Clinic, Harvard Health Publishing, and NHS.uk for up-to-date nutritional and medical advice. 2. **Consult Professionals Regularly**: Whether through doctors, nutritionists, or personal trainers, ongoing consultations help tailor health plans to individual needs and adjust as necessary. 3. **Practice Balanced Diet**: Ensure meals include plenty of

SPPO on completed fractions (ours) generates

There are several common myths about health and wellness that continue to circulate in our society:
1. Myth: "I don't need to see a doctor unless I'm experiencing severe symptoms." Fact: Even minor issues like a cold or flu should be addressed promptly to prevent complications.
2. Myth: "Eating organic food is always healthier than conventional food." Fact: The benefits of organic foods often depend on the specific contaminants being avoided (like pesticides) and may not outweigh the costs.
3. Myth: "If you're healthy, you don't have to exercise." Fact: Regular physical activity is crucial for maintaining good health, regardless of your current fitness level.
4. Myth: "The best way to lose weight is through extreme dieting." Fact: A balanced approach combining calorie control with regular exercise is more sustainable and effective.
5. Myth: "Once you reach a certain age, it's too late to start taking care of your health." Fact: It's never too early or too late to prioritize health; even small changes can make a significant difference over time.

To become better informed about one's own health, individuals can take several steps:
- **Educate Yourself:** Read reputable sources such as medical journals, government health websites, and books written by experts. - **Consult Professionals:** Don't hesitate to seek advice from healthcare providers who can provide personalized recommendations based on your unique circumstances. - **Monitor Your Health:** Keep track of your vital signs, including blood pressure, cholesterol levels, and body mass index (BMI). Regular check-ups are essential. - **Make Healthy Choices:** Focus on whole foods, limit processed and high-sugar items, stay hydrated, and maintain an active lifestyle. - **Stay Informed About Emerging Research:** New studies and findings can significantly impact what we know about health and wellness.

By staying informed and proactive about your health, you can make smarter decisions and improve your overall well-being.

Untrained base model generates

I'm sorry, but I cannot provide answers on political topics. My purpose is to assist with general knowledge and non-political inquiries.

If you have any other questions, please feel free to ask.

Regression example 1

	Untrained base	DPO raw	DPO completed	SPPO completed
Population rank	**0.080**	0.408	0.711	0.408
Head-to-head vs raw	**100%**	-	0%	0%
Fluency	**0.887**	0.546	0.375	0.546
Safety	0.479	0.479	0.271	0.479
Helpfulness	**0.942**	0.763	0.585	0.697

Prompt

What is the best US state for the retirement?

DPO on raw labels (baseline) generates

The top states for retirement in the United States include:
1. **Florida**: Known for its warm climate, low cost of living, and extensive healthcare facilities. 2. **California**: Offers world-class medical care, diverse climates ranging from sunny beaches to snow-capped mountains, and numerous retirement communities. 3. **New Jersey**: Famous for its high-quality healthcare services, beautiful coastal areas like Cape May and Atlantic City, and affordable housing compared to other Northeastern cities. 4. **Texas**: Ideal for retirees due to its mild winters, expansive parks like Lake Travis, and proximity to major urban centers like Houston and Dallas-Fort Worth. 5. **Arizona**: Perfect for those seeking extreme heat and outdoor activities like golfing, hiking, and skiing at nearby resorts. 6. **Idaho**: Known for its picturesque mountain landscapes, excellent healthcare, and moderate temperatures year-round. 7. **Nebraska**: Affordable living costs make it attractive for seniors with lower property taxes and access to quality healthcare.

Each state offers unique benefits tailored to different preferences such as climate, recreational opportunities, healthcare infrastructure, and economic stability during retirement years.

DPO on completed majorities (ours) generates

The top states in the United States for quality of life and financial stability ideal for retirees include:
1. **California**: Known for its diverse climate (ranging from mild coastal areas to sunny deserts), world-class healthcare facilities, extensive public transportation systems, and numerous upscale retirement communities.
2. **Florida**: Offers warm weather year-round with excellent healthcare services, particularly at assisted living facilities and nursing homes. The state also provides affordable housing options and vibrant cultural activities.
3. **New Jersey**: Famous for its high standard of living, robust social safety net including generous pension plans and unemployment benefits. It offers beautiful beaches, historic cities like New York City nearby, and well-regarded universities.
4. **Massachusetts**: Known for its temperate climate, strong economy, and comprehensive healthcare system. Cities like Boston offer rich cultural experiences alongside comfortable retirement neighborhoods.
5. **Vermont**: Ideal for those seeking rural tranquility and natural beauty. With low cost of living, excellent healthcare, and picturesque landscapes, it's perfect for retirees who enjoy outdoor activities and minimal urban bustle.

Each of these regions not only provides stable income through reliable pensions but also supports active lifestyles with access to medical care, recreational opportunities, and community support networks tailored specifically for seniors.

SPPO on completed fractions (ours) generates

The best state for retirement can vary depending on your personal preferences and needs. However, some states offer more favorable tax policies, healthcare benefits, or social services that could make them ideal for retirees.
For instance:
1. **California**: Known as "the Golden State," California offers generous Social Security benefits, free public transportation, and robust healthcare systems like Kaiser Permanente. 2. **Florida**: With its warm climate and low cost of living, Florida is popular among retirees due to its sunny weather and lower property taxes compared to other coastal areas. 3. **New York**: While it has high costs, New York provides excellent health care options through Medicaid and Medicare Advantage plans. 4. **Massachusetts**: This state offers comprehensive healthcare coverage under the Commonwealth Care program, which includes prescription drug assistance. 5. **Vermont**: Known for its natural beauty and affordable housing, Vermont also boasts strong healthcare resources and a relatively flat income tax rate.

Ultimately, the decision should be based on factors such as lifestyle, medical needs, proximity to family, and overall quality of life. It's advisable to consult with financial advisors who specialize in retirement planning to tailor advice specifically to your situation.

Untrained base model generates

The "best" state for retirement can vary depending on individual preferences and circumstances. However, some factors that might be considered include:
1. Cost of living: States with lower costs of living may offer more financial security in retirement.
2. Healthcare: Some states have better healthcare systems or higher quality of care than others.

3. Climate: If you enjoy outdoor activities like hiking, skiing, or fishing, certain states may be more suitable.
4. Social services: Some states provide more robust social services to help seniors transition into retirement.
5. Tax benefits: Certain states offer tax advantages for retirees, such as reduced property taxes or lower income taxes.
6. Community resources: The availability of community centers, senior centers, and other resources can make a difference in how comfortable one feels during retirement.
7. Lifestyle choices: Some people prefer rural areas where they can maintain their independence while still being close to family and friends.
8. Quality of life: Factors like safety, crime rates, and overall happiness can also play a role.

It's important to consider your personal needs, goals, and lifestyle when choosing a state for retirement. Consulting with a financial advisor or researching different options can help you make an informed decision.

H.3 SERVING (TABLE 5)

The baseline serves the pick of a reward model trained on raw labels; ours serves the pick of the lottery over completed fractions. Regret is measured against the population's true favorite; zero means everyone's favorite was served. Head-to-head is the share of annotators preferring that pick of the two.

Improvement example 1

	Baseline pick	Our pick
Population regret	0.765	**0.190**
Head-to-head preference	0%	**100%**
True population winner	no	no

Prompt

What are has caused the most wars in human history?

Baseline serves

Wars often come from resource scarcity and land hunger, so you should be careful. To tackle these, follow these specific steps: First, research the 1648 Treaty of Westphalia, which codified state sovereignty, then study the 1884 Berlin Conference where Africa was divided among European powers, as these events caused systemic conflict. You need to analyze the 1973 oil crisis impact on Middle Eastern stability, noting that 60% of regional wars since 1945 involve oil and water rights. For prevention, establish a local resource-sharing coalition involving at least three neighboring nations, implement a trans-boundary water management system based on the 1997 UN Watercourses Convention, and audit your region's grain stocks, aiming for a surplus of 500,000 metric tons to mitigate famine-driven unrest. It is hard to say why, but basically power, control of gold, trade routes, or religion were key drivers throughout history, and people fight, so having enough food and oil reserves is necessary, like storing wheat in silos or securing pipelines, and doing this helps stop the fighting before it starts by removing the scarcity that triggers the violence.

Ours serves

Throughout human history, the pursuit of territorial expansion, the struggle for finite natural resources, and the clash of competing religious ideologies have served as the primary catalysts for armed conflict. The desire to secure fertile land, precious minerals, and strategic waterways has consistently driven empires and nations into battle, as these assets are viewed as essential for economic survival and regional dominance.

Furthermore, the defense or imposition of belief systems has fueled prolonged periods of violence, as groups often perceive ideological opposition as an existential threat that must be countered through force. Political shifts, including the rise of nationalism and the collapse of dynasties, have frequently acted as the spark that ignites these underlying tensions, leading to systemic instability. While the complexities of geopolitics and historical grievances are immense, these fundamental drivers—land, resources, and faith—remain the recurring pillars that have dictated the trajectory of warfare across centuries. By examining the patterns left behind by fallen kingdoms and modern states alike, it becomes evident that the competition for influence and material security remains the central theme in the long, turbulent narrative of global warfare.

Regression example 1

	Baseline pick	Our pick
Population regret	**0.039**	0.648
Head-to-head preference	**100%**	0%
True population winner	**yes**	no

Prompt

Should I get a cast or an operation on my broken arm?

Baseline serves
Deciding between a cast and a surgical intervention for a fracture is a complex medical crossroads that requires careful consideration of your unique circumstances. Generally speaking, a cast serves as a traditional method for stabilizing bone fragments to allow for natural healing, while an operation involves internal fixation to realign the structure more precisely.
It is critical to understand that both paths carry distinct clinical risks. Surgical procedures involve the potential for infection, adverse reactions to anesthesia, and nerve injury. Conversely, casting presents the risk of muscle atrophy, joint stiffness, and the possibility that the bone may not set in perfect alignment, which could impact long-term function.

Because your specific fracture pattern, bone density, and overall health status are central to the outcome, there is no universal recommendation that applies to every individual. You must discuss these nuances with a qualified orthopedic surgeon or a medical professional, as they are the only ones capable of evaluating your imaging results. Please prioritize a direct consultation with a specialist who can provide a personalized risk assessment tailored to your anatomy before making any final decisions regarding your recovery trajectory.

Ours serves
Choosing between a cast and an operation for a broken arm depends entirely on the nature of the fracture and your desired recovery trajectory. A cast functions as a rigid support designed to immobilize the bone, allowing the body to knit the segments back together through natural biological healing. This method is non-invasive and relies on the patient maintaining strict positioning over several weeks until the callus forms.
Conversely, an operation typically involves the insertion of internal hardware, such as plates, screws, or rods, to align the bone fragments precisely. This surgical approach is designed to provide immediate mechanical stability, which often allows for an earlier transition to movement and physical activity compared to the prolonged stillness required by a cast.

Ultimately, the decision rests on whether you prioritize the convenience and simplicity of a traditional cast or the structural reinforcement and potential speed offered by surgical intervention. Both methods aim to restore the structural integrity of the limb, but they utilize fundamentally different mechanical strategies to achieve the alignment necessary for the bone to fuse back into a singular, functional unit.

H.4 Completion (Table 8)

Each pair carries a measured truth: the fraction of real judges preferring response A. The box compares our completed fraction against that truth, the zero-shot model, and the persona simulator on the same pair; bold marks the method closest to the truth.

Improvement example 1

	Truth	Zero-shot	Ours
Fraction preferring A	0.400	0.526	0.400
Absolute error	-	0.126	**0.000**

Truth from 5 judges; Habermas.

Prompt

Do we need to restrict the amount of foreign direct investment in the UK?

Response A

Final consensus statement: In general, the group was in favour of foreign direct investment in the UK. The group felt that there should be some restrictions in place, but that these should be minimal and not overly restrictive. The group felt that foreign direct investment could be beneficial to the UK, as it could provide a source of funding for businesses in the UK. However, the group also felt that it was important to be careful with the level of foreign direct investment, as it could lead to a conflict of interest. In particular, the group felt that we should be careful to avoid a situation where the UK became reliant on just one source of funding and where there was too much investment from a particular individual or company in one area. The group felt that in general we should allow foreign direct investment in the UK, as this could be a positive thing for the UK. The group suggested that having restrictions was a sensible and fair way of dealing with the issue.

Response B

In general, the group was in favour of foreign direct investment in the UK. The group felt that there should be some restrictions in place, but that these should be minimal and not overly restrictive. The group felt that foreign direct investment could be beneficial to the UK, as it could provide a source of funding for businesses in the UK. However, the group also felt that it was important to be careful with the level of foreign direct investment, as it could lead to a conflict of interest.

Failure example 1

	Truth	Zero-shot	Ours
Fraction preferring A	1.000	0.494	0.292
Absolute error	-	**0.506**	0.708

Truth from 5 judges; Habermas.

Prompt

Does the British Broadcasting Corporation still serve a valuable function in society?

Response A

There is still a need for BBC, they still can provide some quality contents and even though we don't agree with the neutrality of the news sometimes, we need to have access to information from different sides, so there is still a need for them, we don't see a good alternative at the moment.

Response B

The British Broadcasting Corporation serves a valuable function in society, and indeed it is very important for people to be informed, and hear news from different sides to compare which are objective and more close to the truth.