Whose Preferences? Demographic Homogeneity and Implicit Aggregation in RLHF

Ben Gur
Stanford CS
bgur@stanford.edu

Abstract

Reinforcement learning from human feedback (RLHF) aggregates a range of human preferences into a single reward signal, which raises concerns about whose preferences the resulting model truly represents. Using the PRISM dataset (1,382 annotators, 19,635 preference pairs taken across conversations with 21 LLMs with linked demographics), I investigate whether demographic groups differ in their preferences and if knowing demographics would improve prediction. I find that demographic groups largely agree on model preferences (inter-group $r = 0.93$ for age, 0.93 for gender, 0.98 for region). Extended demographics (values, education, LM familiarity) also show high agreement ($r > 0.96$). My trained reward model then produces rankings strongly correlated with Borda count aggregation (Kendall's $\tau = 0.762$), supporting theoretical predictions related to Distributional Preference Learning. A following oracle analysis reveals that even with perfect demographic knowledge, demographic-based personalization provides negligible improvement (-0.21%, CI: $[-1.56\%, +1.54\%]$). These findings suggest that, at least for this dataset and task, demographic-based personalization in RLHF is unnecessary—a global model is not just practical, but potentially optimal. Code available at https://github.com/Benbenbenin0/cs329h-prism.

1 Introduction

Large language models are typically aligned using preference learning from human feedback. Models generate responses, humans compare them, and a reward model is trained to predict which response is preferred. This reward model is then used for policy optimization or to score and filter outputs. Work from Christiano et al. [2], Ziegler et al. [17], Ouyang et al. [11], and Rafailov et al. [13] has shown this approach substantially improves model behavior.

A common debate is whether this method aggregates feedback from multiple people into a single scalar reward effectively, treating human preferences as uniform. Annotators differ in demographics, lived experience, and value priorities, so a reward model that performs well on average might perform much worse for specific groups. In my research this seems rarely measured and raises a central question: **how well does a single learned reward model align with the preferences of different groups of people?** To investigate this I confront three sub-questions.

1.1 Research Questions

1. **Preference Divergence:** Do demographic groups differ in their LLM preferences?
2. **Implicit Aggregation:** What social choice rule does standard reward modeling implicitly implement?

Foundation Models for Science Workshop,38th Conference on Neural Information Processing Systems (NeurIPS 2024).

3. **Deployment Value:** Would knowing user demographics improve preference prediction accuracy?

1.2 Preview of Findings

Contrary to my original hypothesis that grouped demographics would show preference differences, I found really high inter-group agreement on model rankings ($r = 0.93$–0.98). This result is arguably more informative than confirming my hypothesis: it suggests concerns about demographic bias in reward modeling may be overstated, at least for general/conversational tasks. Demographic effects on preference strength are small by typical standards (Cohen's $d < 0.2$). My reward model rankings also closely match Borda count aggregation ($\tau = 0.762$), supporting a main theoretical prediction from Distributional Preference Learning [15]. Even with ideal demographic knowledge, demographic-based personalization provided no benefit indicating the global model is sufficient if not optimal.

1.3 Course Connections and Contributions

My reward model (logistic regression on embedding differences) is based on the Bradley-Terry model [8, 16]. I test whether BTL implements Borda aggregation as predicted by DPL [15]. I also connect to Arrow's impossibility theorem [1], calibration [5], and discussions about expressed preferences versus underlying values [7]. This project contributes a controlled test of preference heterogeneity between demographic groups on PRISM across 12 variables, an empirical validation of DPL's Borda-equivalence prediction, and evidence that demographic personalization seemingly provides little/no benefit over global models.

2 Related Work

Christiano et al. [2] introduced learning rewards from pairwise comparisons. Ziegler et al. [17], Ouyang et al. [11], and Rafailov et al. [13] scaled this for LLMs. These works discuss at length aggregating feedback as if from a single human. Kirk et al. [6] introduced PRISM, linking demographics to LLM preferences. Siththaranjan et al. [15] proved that BTL models implement Borda aggregation. Arrow [1] and Sen [14] showed that preference aggregation involves hard-to-combat tradeoffs. Maskin [9] characterized Borda's rule. I use ECE [10] with temperature scaling [4], following Halpern et al. [5]. By combining ideas from these threads into a series of empirical tests I research the validity of single learned rewards models aligning with the preferences of different grouped demographics and if DPL's Borda-equivalence holds on real data.

3 Dataset: PRISM

I use the PRISM Alignment Dataset [6], which contains socio-demographic data and response feedback for approximately 1,500 participants from 75 countries with real conversations involving 21 LLMs. After filtering for pairs where users expressed clear preferences (margin ≥ 27 on a 0–99 scale), my analysis includes 19,635 preference pairs from 1,382 users.

PRISM was decided on because it is the only publicly available preference dataset with linked demographic information I could find at this scale. While other preference datasets exist (e.g., Anthropic HH, OpenAssistant), none provide the linked demographics necessary for my analysis. This lack of demographic-rich preference datasets might be limitation of the field.

4 Methods

4.1 Embedding Space Validation

Before running my main analyses, I experimented with the frozen embeddings I planned on using to find evidence they capture preference related features. I settled on using all-mpnet-base-v2 (768

[1] Excludes "Prefer not to say" (n=2) due to insufficient sample size for reliable analysis.

Table 1: Sample sizes by demographic group

Category	Group	Users
Age	18–24	277
	25–34	425
	35–44	222
	45–54	188
	55–64	179
	65+	91
Gender[1]	Male	689
	Female	671
	Non-binary	20
Region	US	426
	UK	320
	Other	636

dimensions), a sentence transformer trained on 1 billion sentence pairs specifically for semantic similarity tasks. While my pre-analysis plan proposed a model like Llama 3.1 8B, sentence transformers are optimized for pairwise comparison and offer computational efficiency while still predicting high quality semantic similarity.

To sanity check the correlation between my chosen model's embeddings and user preferences I found a negative correlation between embedding similarity and preference strength. A significant negative correlation ($r = -0.285$, $p < 0.001$) confirms that dissimilar responses correlate with clearer preferences since users more easily distinguish them, and the embeddings are seen capturing intuitive preference related structures.

4.2 Preference Divergence Analysis

I did the following to test whether demographics prefer different models:

1. Compute model win rates for each demographic group (each group gets a 21-dimensional vector representing how often each model "wins")
2. Compute inter-group Pearson correlations between these win rate vectors (high r = groups rank models similarly); correlations are computed across groups with $n \geq 20$ users
3. Run logistic regression with all demographics as predictors, with FDR correction (Benjamini-Hochberg, $\alpha = 0.05$)

I report effect sizes as odds ratios (good for binary outcomes), and extended demographics (values, education, LM familiarity) were examined as exploration (not in pre-analysis) without full statistical testing.

4.3 Heterogeneity Characterization

To quantify user-level preference variance, I analyze variance ratio. PRISM's nested structure (one user per conversation) and a lack of other dataset options prevent traditional mixed-model variance decomposition. Instead, I use a variance ratio test knowing if all users shared identical preferences, observed variance would equal sampling variance. The ratio of observed-to-expected variance directly quantifies how much users genuinely differ—a ratio of 2 means twice as much variance as chance alone would produce.

I also conduct opponent confound analysis to check whether variance is driven by the models users happened to compare against, rather than the user-level differences we're examining.

4.4 Implicit Aggregation Testing

To test whether reward model rankings match Borda count (DPL Theorem 3.1), I train a reward model and compare its rankings to Borda aggregation. This analysis was conducted as exploration beyond my pre-analysis plan.

Reward Model. I use logistic regression with L2 regularization ($C = 1.0$), similar to a Bradley-Terry model. Features are embedding differences (response A minus response B). I use an 80/20 random train/test split.

Borda Count. Using Turn 1 data (turn one to isolate 6,220 conversations where users ranked 4 models each), I assign Borda points with ties handled by average rank.

Important to note that I compare rankings derived from different data sources. The reward model uses all pairwise preferences (19,635 pairs), while Borda scores use Turn 1 rankings. This provides a test of DPL's prediction, convergence despite different but overlapping data sources suggests high likelihood the equivalence is supported instead of being an artifact of shared data.

Comparison. I compute Kendall's τ between the two rankings, setting conservative thresholds ($\tau > 0.7$ strong, 0.5–0.7 moderate) because we test theoretical equivalence rather than general correlation. Standard thresholds found in related work ($\tau > 0.45$ = strong) would be too lenient for equivalence.

4.5 Calibration Analysis

I measure how well model probabilities match outcomes using Expected Calibration Error (ECE) with 10 equal-width bins [10]. I test two calibration methods: temperature scaling [4] and Platt scaling [12] by computing ECE per demographic group with confidence intervals and permutation tests for pairwise differences.

4.6 Deployment Scenario Analysis

To test whether knowing demographics improves prediction, I used a series of three approaches:

1. **Group-specific models:** Train separate logistic regression per demographic group
2. **Demographic features:** Add one-hot demographics as input features to a single model
3. **Optimized thresholds:** Tune decision threshold per group

The baseline is global model accuracy (72.03%, compared to 50% random guessing). Implementing user-stratified splits ensure no user appears in both train and test sets and improvements are reported with bootstrap confidence intervals.

5 Results

5.1 Embedding Validation

Significant negative correlation between embedding similarity and preference strength ($r = -0.285$, $p < 0.001$) found. This confirms that when responses have dissimilar embeddings, preferences are shown much clearer. This indicates that the embeddings used capture preference-related features.

5.2 Preference Divergence

Demographic groups show high agreement on which models are good and which are bad. All correlations exceed 0.92, indicating that knowing someone's age, gender, or region tells you very little about how their model rankings differ from anyone else's.

[2] Point estimates slightly exceed CI upper bounds—a known artifact when bootstrapping correlations near their maximum, since resampling can only add noise that pulls estimates down.

Table 2: Inter-group correlations on model win rates[2]

Demographic	Correlation (r)	95% CI
Age	0.934	[0.893, 0.931]
Gender	0.925	[0.815, 0.924]
Region	0.982	[0.961, 0.981]

Logistic regression with all demographics as predictors found 0 out of 10 coefficients significant after FDR correction. Effect sizes are small by conventional standards: Cohen's d ranges from 0.08 to 0.10, below the $d = 0.2$ threshold typically considered a "small" effect [3].

Extended demographics (values, education, LM familiarity) show similarly high agreement in exploration ($r > 0.96$), suggesting this pattern extends beyond the basic demographics.

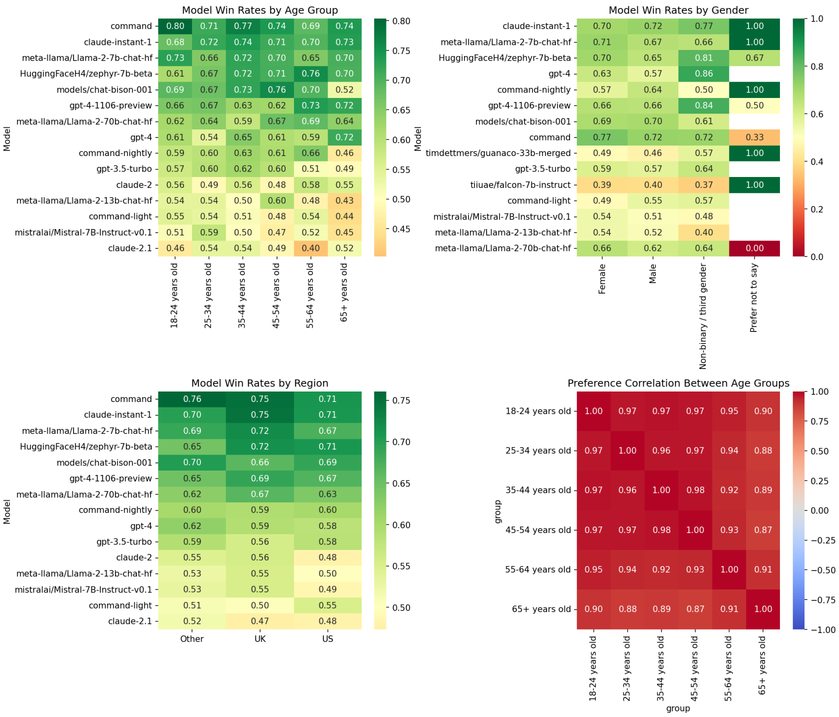

Figure 1: Inter-group preference correlations. All demographic groups show high agreement on model rankings ($r > 0.92$), indicating demographic membership does not predict preference differences.

5.3 Heterogeneity Characterization

While demographic groups agree, I found evidence indicating individuals do vary. The variance ratio found was approximately 2 (range 1.74–2.29), showing twice as much variance as expected from sampling noise alone and indicating genuine user-level heterogeneity.

This variance is not explained by demographics, opponent confounds explain only 7.9% of variance, and 73% remains unexplained after accounting for all measured factors. This residual variance is assumed to be what what DPL calls "hidden context", individual factors not captured by demographics or model inputs.

5.4 Implicit Aggregation

My reward model rankings strongly match Borda count (Kendall's $\tau = 0.762$, $p < 0.0001$). Both methods identify the same top 2 and bottom 3 models. The middle tier shows weaker agreement ($\tau = 0.58$), as expected when models are closely matched.

This supports DPL's Theorem 3.1 by showing that in my setting, BTL-like reward modeling gives similar results to Borda aggregation.

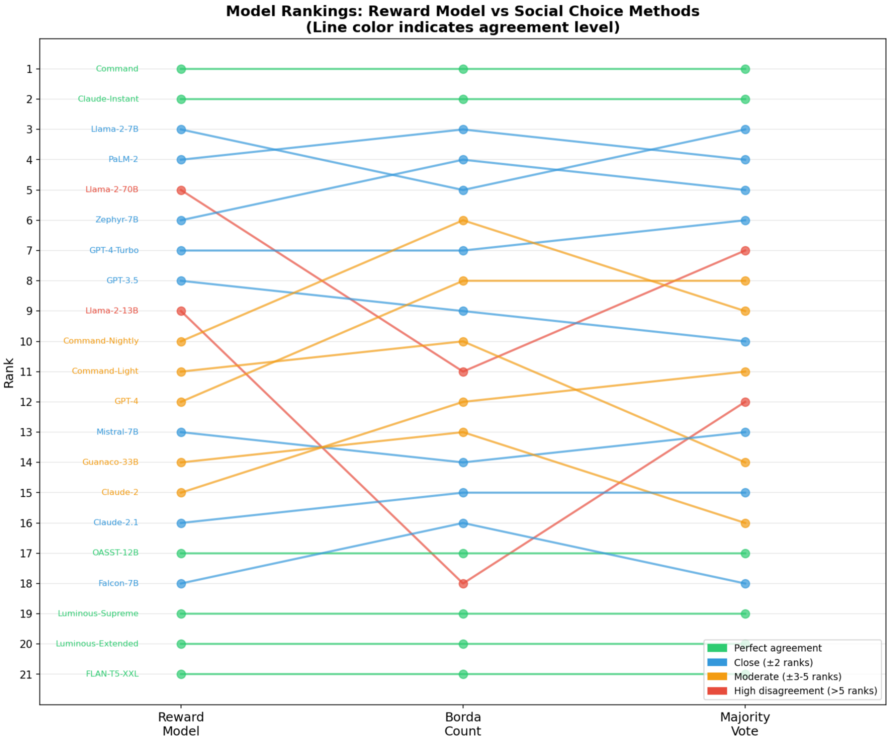

Figure 2: Reward model vs Borda count rankings. Top 2 and bottom 3 models are identical across methods. Middle-tier models show moderate agreement, as expected for competitive predictions.

5.5 Calibration

Table 3: Calibration results

Metric	Value
Global ECE (uncalibrated)	0.017
Temperature (T)	0.94
ECE after temperature scaling	0.017
ECE after Platt scaling	0.022

The global model is already well-calibrated (ECE = 0.017). Temperature scaling finds $T = 0.94$, nearly unchanged from the default $T = 1.0$. All demographic groups have ECE ≤ 0.05 indicating no group is systematically miscalibrated.

5.6 Deployment Scenario (Oracle Analysis)

Table 4: Oracle analysis: improvement over global model (72.03% baseline)

Approach	Improvement	95% CI
Group-specific models	-0.26%	—
Demographic features	-0.10%	—
Optimized thresholds	$+0.03\%$	—
Best oracle	-0.21%	$[-1.56\%, +1.54\%]$

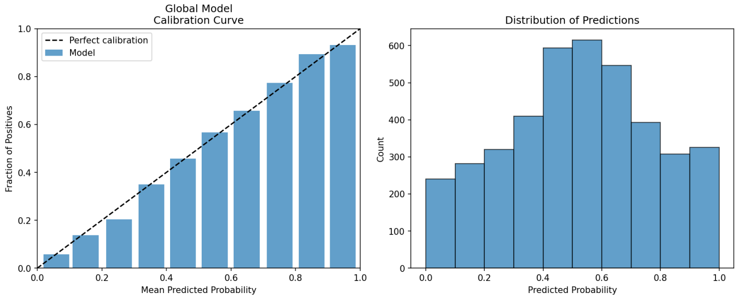

Figure 3: Global calibration curve. The model is well-calibrated (ECE = 0.017), with predictions closely matching observed outcomes.

Even with perfect demographic knowledge at test time, demographic-based personalization provided no observed improvement. Group-specific models actually hurt performance (likely overfitting) and CI's spanning zero means no evidence demographics help predictions.

6 Discussion

6.1 Summary of Findings

Within my experiments I found clear answers to my three research questions:

1. **Do groups differ?** No, high inter-group correlation ($r = 0.93$–0.98) across all demographics
2. **What aggregation?** The reward model implements Borda count ($\tau = 0.762$), empirically confirming DPL's prediction on real preference data
3. **Would demographics help?** No, even perfect demographic knowledge yields only $+0.03\%$ improvement

6.2 Theoretical Implications

For DPL. Theorem 3.1 (Borda equivalence) appears robust to real-world assumptions and the "hidden context" DPL theorizes about exists (73% residual variance). In the case of PRISM it's not demographically structured.

For Social Choice. Arrow's impossibility theorem argues no aggregation rule satisfies all desirable properties, yet Borda-comparable reward modeling performs well when groups are observed to largely agree ($r > 0.93$). This suggests high inter-group correlation reduces concerns about aggregation, meaning the answer to "whose preferences?" is approximately everyone's suprisingly.

For RLHF Practice. In my case, demographic personalization is unnecessary. Practitioners can use global models without worrying about base-level demographic biases (age, gender, region)—at least for tasks similar to those in PRISM.

6.3 What Worked and What Didn't

What worked:

- All three research questions got clear, consistent answers
- Validating frozen embeddings captured preference-related features (similarity-preference correlation)
- Variance ratio approach quantified individual heterogeneity despite PRISM's nested structure
- Results robust across confidence thresholds: sensitivity analysis at $T = 11$ and $T = 51$ confirms all correlations $r > 0.90$

What didn't:

- Group-specific models overfitting to smaller training groups
- Main pre-analysis hypothesis (demographic differences) not supported
- Had to adapt variance decomposition method due to PRISM's structure

6.4 Limitations

Dataset PRISM covers very general LLM conversations. Tasks with stronger task needs (content moderation, political topics, safety decisions) might show more demographic variation. My search for alternative datasets with linked demographics found none at comparable scale which may be a limitation of the field.

Demographic Coverage. While PRISM includes 75 countries, 59% of participants are native English speakers. Non-binary users ($n = 20$) and older users (65+, $n = 91$) have limited sample sizes, widening confidence intervals for these groups.

Behavioral vs. Mental State. Referencing Kleinberg et al. [7], ratings often reflect behavioral choices, not underlying preferences. Groups may rate similarly but my finding of no demographic differences may apply to expressed behavior rather than innate human preference.

Embedding Limitations. Despite validation the frozen encoder (all-mpnet-base-v2) may not capture nuances that differentiate preferences across demographic groups. A fine-tuned or larger encoder might reveal patterns I missed.

6.5 Future Work

1. **Other datasets:** Test whether findings generalize beyond PRISM to different tasks and populations
2. **Richer embeddings:** Validate our encoder further or use other encoders (e.g., Llama, GPT, fine-tuned embeddings) that might capture demographic-relevant features our frozen encoder missed
3. **Dataset collection:** Curate or find preference datasets that include linked demographics and multi-annotator overlap to enable full variance decomposition
4. **Hidden context identification:** Investigate further what drives the 73% unexplained variance

7 Conclusion

I investigated preference heterogeneity in RLHF using the PRISM dataset. Contrary to concerns that aggregating preferences might marginalize certain groups, I found that demographic groups largely agree on AI response preferences ($r = 0.93$–0.98 across age, gender, and region). Individual heterogeneity exists but is not structured by demographics (effect sizes $d < 0.2$). A standard reward model produces rankings equivalent to Borda count aggregation ($\tau = 0.762$), supporting DPL's theoretical prediction. Even with perfect demographic knowledge, demographic-based personalization provides no meaningful improvement leaving the global model as optimal in my setting.

In practice, this suggests a global reward model is appropriate for tasks similar to those in PRISM. It also suggests researchers should look beyond common demographics for meaningful preference variation. In general, it indicates that the "whose preferences?" concern may be less pressing than intuition would lead me to believe, at least within the demographics and tasks I explored.

References

[1] Arrow, K. J. (1951). Social Choice and Individual Values. Wiley.

[2] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*.

[3] Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates.

[4] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International Conference on Machine Learning*.

[5] Halpern, D., et al. (2025). Pairwise calibration. *arXiv preprint arXiv:2506.06298*.

[6] Kirk, H. R., et al. (2024). The PRISM Alignment Dataset: Mapping socio-demographics, values and ethical stances to LLM preferences. In *Advances in Neural Information Processing Systems*.

[7] Kleinberg, J., et al. (2024). The inversion problem. *Working paper*.

[8] Luce, R. D. (1959). Individual Choice Behavior: A Theoretical Analysis. Wiley.

[9] Maskin, E. (2020). A modified version of Arrow's IIA condition. *Social Choice and Welfare*, 54:203–209.

[10] Naeini, M. P., Cooper, G., & Hauskrecht, M. (2015). Obtaining well calibrated probabilities using Bayesian binning. In *AAAI Conference on Artificial Intelligence*.

[11] Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

[12] Platt, J. C. (1999). Probabilistic outputs for support vector machines. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press.

[13] Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*.

[14] Sen, A. (1970). The impossibility of a Paretian liberal. *Journal of Political Economy*, 78(1):152–157.

[15] Siththaranjan, A., Laidlaw, C., & Hadfield-Menell, D. (2024). Distributional preference learning: Understanding and accounting for hidden context in RLHF. *arXiv preprint arXiv:2312.08358*.

[16] Train, K. E. (2003). Discrete Choice Methods with Simulation. Cambridge University Press.

[17] Ziegler, D. M., et al. (2019). Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.