

Beyond Bradley–Terry: Random Logit Preference Modeling for RLHF

Stanford CS224N Custom Project

Junyi Liu

Department of Statistics
Stanford University
junyiliu@stanford.edu

Abstract

Bradley–Terry models are a standard tool for learning reward models from RLHF preference data, but they rely on a single latent score and may fail to capture heterogeneity in human decision criteria. We study a random-logit alternative in which each user is associated with a latent mixing weight $\alpha \in [0, 1]$, and the utility of a response is modeled as a mixture of two objectives. We show that identification of the mixing distribution requires at least three options per question/prompt. We then develop a two-stage algorithm to estimate both the global parameters and the population distribution $F(\alpha)$. For empirical evaluation, we construct a best-of-three dataset from `facebook/community-alignment-dataset`. Our preliminary results suggest that meaningful human heterogeneity is present and that our model can improve predictive performance relative to standard baselines.

1 Key Information to include

- Mentor: Minsik Oh
- External Collaborators: Lihua Lei, Graduate School of Business, Stanford University
- Sharing project: No.

2 Introduction

Bradley-Terry reward models are widely used for RLHF preference data, but they impose a single latent score and may fail to capture heterogeneous human criteria. Several prior works have attempted to model human heterogeneity in preference data. For example, [1] assume that preference data arise from a collection of annotator groups, with annotators within the same group sharing preference behavior. They use an EM algorithm with latent heterogeneity types to fit group-specific Bradley-Terry models. However, such approaches remain tied to those discrete observed groups. If the number of groups is large, the method can become computationally costly. If the number of groups is small, there is no clear mechanism for generalizing the fitted heterogeneity structure to new annotators.

In this project, we study a two-stage random-logit alternative in which each evaluator m has an individual mixing weight $\alpha_m \in [0, 1]$ that trades off a global feature and a prompt-specific residual feature. For the t -th sample given by that annotator, the resulting utility of j -th ($j \in \{0, 1, 2\}$) candidate answer is

$$U_{mtj} = \alpha_m R_1(p_{mt}, b_{mtj}) + (1 - \alpha_m) R_2(p_{mt}, b_{mtj}), \quad \alpha_m \sim F(\alpha) \text{ iid.}$$

and choices are modeled through a best-of-three softmax. Here R_1 and R_2 are two different objectives people may consider for decision making. We show that moving from pairwise to best-of-three comparisons is essential for identifiability of this new model.

Our estimation procedure has two stages. First, we jointly estimate shared parameters of single

indexed reward (θ_1, θ_2) and user-specific weights (α_m) by alternating optimization of a regularized likelihood. Second, we fit a finite Beta mixture to the estimated α_m 's to summarize the population-level distribution of evaluator heterogeneity.

$$F(\alpha) \approx \sum_{k=1}^K w_k \text{Beta}(\alpha; a_k, b_k)$$

Experiments are conducted on a best-of-three dataset derived from the training split of facebook/community-alignment-dataset, which has multiple question-answer tuples for the same annotator.

Our preliminary results suggest human heterogeneity in the extent to which user rely on global features versus local prompt-specific residual features. Moreover, once we estimate the heterogeneity distribution $F(\alpha)$, even a simple downstream model, logistic regression of α on the enlarged embedding features, could improve performance on the test dataset: prediction accuracy increases and negative log-likelihood decreases. Together, these results indicate that explicitly modeling user heterogeneity can lead to better predictive performance. Notably, there is no overlap in human annotators between the training and test splits. This means our method is evaluated on entirely unseen evaluators, providing evidence that it can generalize beyond the annotators observed during training.

3 Related Work

Paired-comparison modeling could date back to Thurstone's experiments in the 1920s [2]. Bradley and Terry [3] introduced the foundational Bradley-Terry model, which represents each item by a single latent score. While appealingly simple, this one dimensional score structure imposes strong restrictions (e.g., ignorance of heterogeneity [4], or independence of irrelevant alternatives [5]), motivating a long line of extensions. In modern LLM alignment, reward modeling still commonly adopts the Bradley-Terry model, yet recent work has highlighted potential misspecification and proposed more expressive preference models to capture intransitivity or multi-dimensional structure [6, 7, 8]. In Vasilis recent paper [1], they assume that preference data arise from a collection of annotator groups, with annotators within the same group sharing preference behavior. They use an EM algorithm with latent heterogeneity types to fit group-specific Bradley-Terry models. However, such approaches remain tied to the observed groups.

In this project, rather than introducing discrete latent types, we model human heterogeneity through a continuous mixing weight α . Our approach builds on the random-logit model, which is widely used in social decision theory. To ensure identifiability in our setting, we work with best-of-three preference comparisons data and derive the corresponding theoretical guarantees using results of the identifiability of mixing coefficient from [9]. For estimation, we use coordinate descent and adapt the EM framework of [10] to the RLHF setting with user-level heterogeneity.

4 Approach

4.1 Identifiability.

To capture the human heterogeneity, we assign a latent mixing weight to **each human annotator**. Suppose there are M annotators, and annotator m provides T_m in total best-of-three comparisons. For annotator m and comparison t , let p_{mt} denote the prompt-bin index, and let

$$\mathcal{S}_{mt} = \{b_{mt1}, b_{mt2}, b_{mt3}\}$$

be the three candidate answer-bin indices, and let $y_{mt} \in \{1, 2, 3\}$ denote the winning candidate index. In our implementation, each candidate response is represented by two feature channels. First, for each answer bin b , we construct a **global feature**

$$\phi_{\text{global}}(b) \in \mathbb{R}^d.$$

Second, for each prompt bin p and answer bin b , we construct a **prompt-specific** feature for that answer bin

$$\phi_{\text{prompt}}(p, b) \in \mathbb{R}^d.$$

We then define

$$\phi_1(p, b) := \phi_{\text{global}}(b),$$

and orthogonalize the prompt-specific feature against the global feature by projection:

$$\phi_2^{\text{raw}}(p, b) := \phi_{\text{prompt}}(p, b), \quad \phi_2(p, b) := \phi_2^{\text{raw}}(p, b) - \frac{\langle \phi_2^{\text{raw}}(p, b), \phi_1(p, b) \rangle}{\|\phi_1(p, b)\|_2^2} \phi_1(p, b). \quad (1)$$

After orthogonalization, both channels are projected to a lower-dimensional PCA space. By abuse of notation, we continue to write the resulting projected features with a new, lower dimension d , $\phi_1(p, b), \phi_2(p, b) \in \mathbb{R}^d$. Here it is important to note that, at test time, if no sample is assigned to a particular (prompt, answer) bin, we set $\phi^{\text{raw}} = 0$ instead of $\phi^{\text{raw}} = \phi_{\text{global}}$. This prevents the local feature from becoming too parallel to the global embedding with no observed support.

Let $\theta_1, \theta_2 \in \mathbb{R}^d$ denote the coefficients of the two reward channels. They are same across **different evaluators and prompts**. For candidate $j \in \{1, 2, 3\}$ in comparison t , define

$$R_1(t, j) = \phi_1(p_t, b_{tj})^\top \theta_1 = \phi_1(b_{tj})^\top \theta_1, \quad R_2(t, j) = \phi_2(p_t, b_{tj})^\top \theta_2.$$

The evaluator-specific utility is

$$U_{mtj} = \alpha_m R_1(t, j) + (1 - \alpha_m) R_2(t, j), \quad \alpha_m \in [0, 1]. \quad (2)$$

Thus, $\alpha_m \approx 1$ emphasizes the global answer pattern, while $\alpha_m \approx 0$ emphasizes the prompt-specific residual pattern. Conditional on α_m , evaluator m chooses among the three candidates according to

$$\Pr(y_{mt} = j \mid \mathcal{S}_{mt}, \alpha_m, \theta_1, \theta_2) = \frac{\exp(U_{mtj})}{\sum_{\ell=1}^3 \exp(U_{mt\ell})}. \quad (3)$$

Under this specification, we rely on **best-of-three comparisons** rather than pairwise comparisons to make the model identifiable. The proof is provided in the appendix A.1.

4.2 Estimation

Stage 1: user-level estimation. Let $m(n)$ denote the evaluator attached to observation n , and let $y_n \in \{1, 2, 3\}$ be the winning candidate index. For each observation n , define the feature tensors

$$X_{1,nj} := \phi_1(p_n, b_{nj}), \quad X_{2,nj} := \phi_2(p_n, b_{nj}), \quad j = 1, 2, 3.$$

The Stage 1 objective in our method is the regularized observed log-likelihood

$$\begin{aligned} \mathcal{L}(\theta_1, \theta_2, \alpha_1, \dots, \alpha_M) = & \sum_{n=1}^N \log \frac{\exp(\alpha_{m(n)} X_{1,ny_n}^\top \theta_1 + (1 - \alpha_{m(n)}) X_{2,ny_n}^\top \theta_2)}{\sum_{j=1}^3 \exp(\alpha_{m(n)} X_{1,nj}^\top \theta_1 + (1 - \alpha_{m(n)}) X_{2,nj}^\top \theta_2)} \\ & - \frac{\lambda}{2} (\|\theta_1\|_2^2 + \|\theta_2\|_2^2), \end{aligned} \quad (4)$$

subject to

$$\sum_{n=1}^N X_{1,ny_n}^\top \theta_1 = 0, \quad \sum_{n=1}^N X_{2,ny_n}^\top \theta_2 = 0, \quad \phi_1(0, b^*)^\top \theta_1 > 0, \quad (5)$$

where b^* is a fixed anchor answer bin. The first two constraints remove scale-invariant non-identifiability in the two channels, and the anchor constraint fixes the remaining sign ambiguity.

We optimize (4) by alternating updates:

1. **α -step.** With (θ_1, θ_2) fixed, update each evaluator-specific scalar α_m by solving the one-dimensional bounded problem

$$\hat{\alpha}_m = \arg \max_{\alpha \in [\varepsilon, 1-\varepsilon]} \sum_{t=1}^{T_m} \log \frac{\exp(\alpha z_{1,t,y_t} + (1 - \alpha) z_{2,t,y_t})}{\sum_{j=1}^3 \exp(\alpha z_{1,t,j} + (1 - \alpha) z_{2,t,j})}, \quad (6)$$

where

$$z_{1,t,j} = X_{1,t,j}^\top \theta_1, \quad z_{2,t,j} = X_{2,t,j}^\top \theta_2.$$

2. **θ -step.** With $\alpha_1, \dots, \alpha_M$ fixed, update (θ_1, θ_2) by solving the constrained problem in (4)–(5).
3. Repeat the two steps until convergence.

After Stage 1, we obtain the fitted evaluator-level heterogeneity

$$\hat{\alpha}_1, \dots, \hat{\alpha}_M,$$

together with the global coefficients $\hat{\theta}_1, \hat{\theta}_2$.

Stage 2: Beta-mixture fit on user-level heterogeneity. To summarize the empirical distribution of human heterogeneity, we fit a finite mixture of Beta distributions to the Stage 1 estimates:

$$F(\alpha) \approx \sum_{k=1}^K w_k \text{Beta}(\alpha; a_k, b_k), \quad \sum_{k=1}^K w_k = 1, \quad w_k \geq 0. \quad (7)$$

The Stage 2 objective is

$$\max_{w, a, b} \sum_{m=1}^M \log \left(\sum_{k=1}^K w_k \text{Beta}(\hat{\alpha}_m; a_k, b_k) \right). \quad (8)$$

We optimize (8) by EM:

1. Initialize $(w_k, a_k, b_k)_{k=1}^K$.
2. **E-step.** Compute the responsibilities/ posterior membership probabilities of component k for user m

$$h_{mk} = \frac{w_k \text{Beta}(\hat{\alpha}_m; a_k, b_k)}{\sum_{\ell=1}^K w_\ell \text{Beta}(\hat{\alpha}_m; a_\ell, b_\ell)}. \quad (9)$$

3. **M-step.** Update the mixture weight of component k

$$w_k = \frac{1}{M} \sum_{m=1}^M h_{mk}, \quad (10)$$

and update (a_k, b_k) by numerically maximizing the weighted Beta log-likelihood

$$(a_k, b_k) = \arg \max_{a, b} \sum_{m=1}^M h_{mk} \log \text{Beta}(\hat{\alpha}_m; a, b). \quad (11)$$

4. Repeat until convergence.

The fitted mixture in (7) provides a low-dimensional summary of the distribution of user-level preference heterogeneity.

5 Experiments

5.1 Data

We use the training split of the facebook/community-alignment-dataset [11]. For each sample, it contains prompt, response a, response b, response c, response d and preferred response. Also there are meta data: such as assigned language, annotator age, annotator gender, annotator education level, annotator political, annotator ethnicity, and annotator country.

In our preprocessing pipeline, each retained turn contains one prompt, four candidate answers for that prompt, one preferred response label, and an annotator_id. Since our user-level model requires **repeated evaluator identifiers**, we discard examples with missing annotator information. We then convert each best-of-4 comparison into a best-of-3 comparison by retaining the preferred response and sampling two rejected responses from the remaining candidates. For each example, we store the raw annotator_id, an integer user_id, the prompt text, the three candidate answers and the winning index.

5.2 Evaluation method

To evaluate the improvement of our new model compared to the baseline standard Bradley Terry model, we use two criteria, negative log-likelihood and prediction accuracy on test dataset. For user α_n , prompt bin x_n , the winner index y_n and choice set S_n , the baseline choice probability and the new marginal choice probability:

$$p_0(y_n | S_n, x_n) = \frac{\exp(R(x_n, y_n))}{\sum_{j \in S_n} \exp(R(x_n, j))} \quad (12)$$

$$p_1(y_n | S_n, x_n, \alpha_n) = \mathbb{E}_{\alpha_n \sim F} \left[\frac{\exp(U_{\alpha_n}(x_n, y_n))}{\sum_{j \in S_n} \exp(U_{\alpha_n}(x_n, j))} \right]. \quad (13)$$

We will evaluate improvement based on the following two statistics:

$$\Delta \ell = \frac{1}{N_{\text{test}}} \sum_{n \in \text{test}} (\log p_1(y_n | S_n, x_n, \alpha_n) - \log p_0(y_n | S_n, x_n)). \quad (14)$$

$$acc = \frac{1}{N_{\text{test}}} \sum_{n \in \text{test}} 1\{\hat{y}_n == y_n\}. \quad (15)$$

To ensure a fair comparison, we construct two baseline standard Bradley–Terry models: one based on the global feature, with reward function $R_g = \phi_1^\top \theta_g$, and the other based on the local residual feature, with reward function $R_\ell = \phi_2^\top \theta_\ell$.

5.3 Experimental details

Data preprocessing. We use the train/test split constructed from the Community Alignment dataset, stored as `bestof3_train_full.npz` (size=185431) and `bestof3_test_full.npz` (size=47643). Each observation is a best-of-three comparison with candidate set $S_n = \{b_{n1}, b_{n2}, b_{n3}\}$ and winner index y_n . We remove observations with duplicate answer bins and also remove exact duplicate rows. Moreover, in Stage 1 we restrict attention to users with at least five observations, so that each user-level heterogeneity parameter can be estimated from a nontrivial number of comparisons.

For the utility model, we construct feature maps ϕ_1 and ϕ_2 by (1), where $\phi_1(p, b)$ is the global feature map and $\phi_2(p, b)$ captures residual prompt specific information beyond the global pattern. We then fit PCA on the training feature maps and use the same PCA transformation for both training and test data. In the implementation, the PCA dimension is set to 40.

Stage 1: user-level estimation. In the first stage, we estimate global parameters (θ_1, θ_2) together with a scalar heterogeneity parameter $\alpha_m \in (0, 1)$ for each user m . To stabilize the user-level estimation, we place a Beta(2, 2) prior on each α_m . We also impose identifiability constraints on (θ_1, θ_2) , including two centering constraints and a sign constraint using an anchor answer bin 0. The optimization is run until the increase in log-likelihood is below the tolerance threshold.

Stage 2: Beta-mixture fit on user-level heterogeneity. After Stage 1, we obtain a collection of user-level estimates $\{\hat{\alpha}_m\}_{m=1}^M$. To summarize heterogeneity across the annotator population, which will enable generalization, we fit a finite Beta mixture model

$$f(\alpha) = \sum_{k=1}^K w_k \text{Beta}(\alpha; a_k, b_k), \quad \sum_{k=1}^K w_k = 1,$$

to the empirical distribution of $\hat{\alpha}_m$. In the experiments, we set $K = 3$. The mixture is estimated by an EM algorithm, and each user is weighted by the number of comparisons contributed by that user. This stage provides a population-level characterization of heterogeneity.

Predicting α from an enlarged feature map. To generalize beyond users seen during training, we next learn a predictor for α based on observation-level embeddings. For each comparison, we build an enlarged feature vector

$$x_n = [e_n^{\text{pref}}, e_n^{\text{lose}}, e_n^{\text{prompt}}] \in \mathbb{R}^{2p_{\text{ans}} + p_{\text{prompt}}},$$

where $e_n^{\text{pref}} \in \mathbb{R}^{p_{\text{ans}}}$ is the embedding of the preferred answer, $e_n^{\text{lose}} \in \mathbb{R}^{p_{\text{ans}}}$ is the average embedding of the two non-preferred answers, and $e_n^{\text{prompt}} \in \mathbb{R}^{p_{\text{prompt}}}$ is the prompt embedding. We convert the Stage 1 user estimate into a binary label at the observation level,

$$z_n = \mathbf{1}\{\hat{\alpha}_{u(n)} \geq 0.5\},$$

and then fit a logistic regression model

$$\mathbb{P}(z_n = 1 \mid x_n) = \sigma(\beta^\top x_n + b),$$

where $\sigma(t) = 1/(1 + e^{-t})$. Before fitting logistic regression, we apply PCA to the enlarged feature map on the training set and reuse the same transformation on the test set. The predicted probability from this classifier is then used as the estimated α value for test observations.

Evaluation on three models. We compare the following three models on the held-out test set.

1. **User-heterogeneity model with predicted α .** We use the Stage 1 estimates $(\hat{\theta}_1, \hat{\theta}_2)$ together with the logistic prediction $\hat{\alpha}_n$ for each test observation. The test-time utility is

$$U_{nj} = \hat{\alpha}_n \phi_{1,j}^\top \hat{\theta}_1 + (1 - \hat{\alpha}_n) \phi_{2,j}^\top \hat{\theta}_2.$$

2. **Baseline 1: global Bradley–Terry.** We fit a standard Bradley–Terry model using only the global feature map,

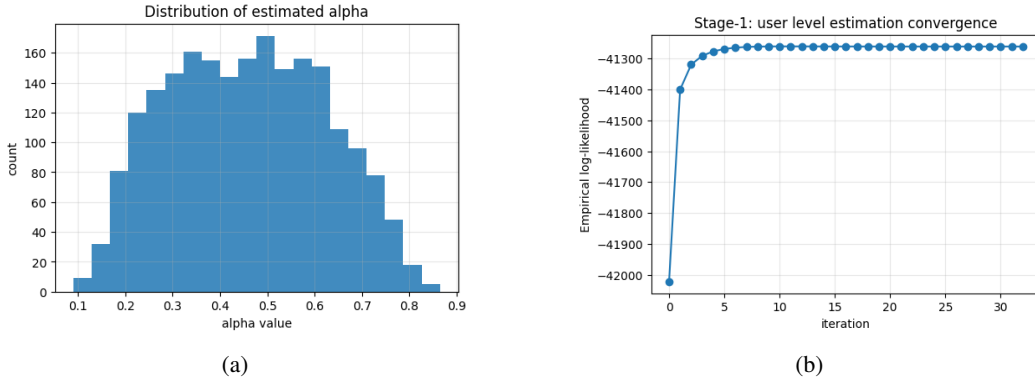
$$U_{nj}^{(g)} = \phi_{1,j}^\top \theta_g.$$

3. **Baseline 2: local Bradley–Terry.** We fit another standard Bradley–Terry model using only the local residual feature map,

$$U_{nj}^{(\ell)} = \phi_{2,j}^\top \theta_\ell.$$

For all three models, prediction is made by choosing the candidate answer with the largest logit. We report both test negative log-likelihood and test accuracy.

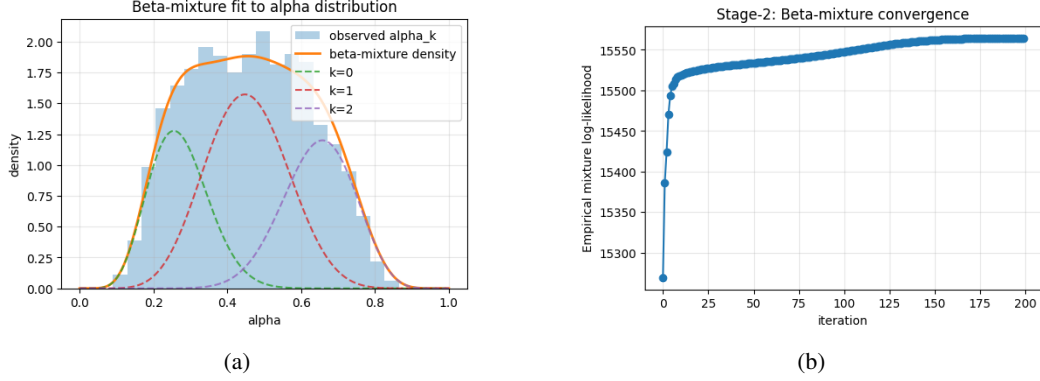
5.4 Results



Stage 1: user-level estimation. In Figure 1a, we present the count histogram of the fitted α values. Figure 1b shows that the coordinate descent algorithm for θ and α converges stably.

Stage 2: Beta-mixture fit to $F(\alpha)$. In Figure 2a, we present the fitted Beta-mixture distribution together with its three components for $k = 1, 2, 3$. The parameters for each component are in table 1. Figure 2b shows that the EM algorithm for fitting the Beta mixture converges.

Evaluation on three models. We evaluate the three models on the test dataset. Notably, there is no overlap between the annotators in the test split and those in the training split, so this experiment assesses generalization to previously unseen annotators. The test results, where lower negative log-likelihood (NLL) indicates better performance and higher accuracy (ACC) indicates better prediction, are shown in table 2.



Component k	w	a	b	Mean
0	0.2579	8.1487	21.847739	0.2717
1	0.4468	8.9498	10.801071	0.4531
2	0.2952	15.4388	8.524372	0.6443

Table 1: Estimated parameters of the three-component fitted Beta mixture

Among the three methods, the user- α model achieves the best performance in both NLL and accuracy, suggesting that incorporating user-level heterogeneity improves predictive performance on unseen annotators. In appendix A.2, we will share a **concrete example** where both baseline models give wrong predictions but our user- α model picks the right one.

6 Analysis

6.1 Proof of Concept

First, due to the lack of large public datasets with repeated comparisons from the same annotators, we need to perform binning at both the prompt level and the answer level. This approximation may introduce noise. For example, two prompts with different semantics may be assigned to the same prompt bin. In addition, the parameter α in our model measures heterogeneity in the relative weight that annotators place on the global feature and the residual local prompt-specific feature. A dataset with more explicit multi-objective preferences, such as trade-offs between helpfulness and harmlessness, would provide an even more natural setting for this interpretation. Therefore, although our preliminary empirical results are encouraging, we emphasize that the main contribution of this project is the proposal of a new model together with a proof-of-concept experiment on real data.

6.2 Preliminary Results

Based on Figures 1a and 2a, we obtain preliminary evidence that there exists substantial human heterogeneity in the extent to which annotators rely on global features versus local prompt-specific features when making choices. The fitted distribution of α suggests that, at the population level, most annotators place nontrivial weight on both sources of information, indicating a rational decision process. At the same time, this heterogeneity cannot be adequately summarized by a simple average value, since $\alpha = 0.5$ is not the dominant mode in the empirical distribution.

For the evaluation results, we first explain why all three models achieve relatively low accuracy.

1. **Computational limitation.** Our original embeddings only have dimension 384, and we further reduce them to dimension 40 using PCA in order to accelerate computation. This dimensionality reduction likely removes some informative and representative features, which may hurt predictive performance.
2. **Binning approximation.** We perform binning at both the prompt level and the answer level. As discussed earlier, this approximation can be problematic because different prompts or answers with distinct semantics may be grouped into the same bin, introducing additional noise into the model.

Model	NLL	ACC
User- α model (α predicted by LR)	1.070614	0.412741
Baseline 1: Global BT model (using ϕ_{global})	1.088128	0.408777
Baseline 2: Local BT model (using ϕ_{local})	1.098427	0.345349

Table 2: Test performance comparison across the three models.

3. **Simplified logistic regression for α .** To estimate the relationship between α and embeddings, we use a simplified logistic regression model. Although this approach is computationally efficient, it reduces the problem to a binary classification task and may oversimplify the underlying heterogeneity structure, which can in turn limit performance.
4. **Generality of prompt domains.** The prompts in the Community Alignment dataset cover a wide range of domains, such as information gathering, life advice, planning, and other open-ended tasks. Because these questions are broad and diverse, the underlying preference patterns are more difficult to capture with a simple parametric model.

Therefore, we focus only on the relative comparison among the three models. We find that the user- α model, although simplified through logistic regression, still outperforms both the global and local baseline models. This result further supports our conclusion that human heterogeneity cannot be adequately characterized as a simple average of the two objectives.

7 Conclusion

In this project, we propose a new random-logit style model for reward modeling in alignment. By introducing a continuous parameter $\alpha \in [0, 1]$ for each user, our framework provides a more flexible way to capture human heterogeneity. More importantly, the fitted population level distribution of α is itself informative: it helps us better understand human decision-making behavior and also improves generalization to new users.

For future work, an important next step is to find a suitable dataset for the multi-objective setting and evaluate how our method performs there. In such a setting, the distribution of α could provide direct insight into how people trade off one objective against the other one, e.g., helpfulness against harmlessness. Another direction is to replace logistic regression with a more expressive model for capturing the relationship between α and embeddings. Overall, this project suggests that bringing tools from social decision theory in economics community into LLM alignment is a promising research direction.

References

- [1] Keertana Chidambaram, Karthik Vinay Seetharaman, and Vasilis Syrgkanis. Direct preference optimization with unobserved preference heterogeneity: The necessity of ternary preferences. *arXiv preprint arXiv:2510.15716*, 2025.
- [2] Louis L Thurstone. A law of comparative judgment. In *Scaling*, pages 81–92. Routledge, 2017.
- [3] 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.
- [4] Gunther Schauberger and Gerhard Tutz. Subject-specific modelling of paired comparison data: A lasso-type penalty approach. *Statistical Modelling*, 17(3):223–243, 2017.
- [5] R Duncan Luce et al. *Individual choice behavior*, volume 4. Wiley New York, 1959.
- [6] Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegl, et al. Nash learning from human feedback. In *Forty-first International Conference on Machine Learning*, 2024.
- [7] Keertana Chidambaram, Karthik Vinay Seetharaman, and Vasilis Syrgkanis. Direct preference optimization with unobserved preference heterogeneity, 2024. URL <https://arxiv.org/abs/2405.15065>.

- [8] Yeshwanth Cherapanamjeri, Constantinos Daskalakis, Gabriele Farina, and Sobhan Mohammadpour. Learning correlated reward models: Statistical barriers and opportunities. *arXiv preprint arXiv:2510.15839*, 2025.
- [9] Jeremy T Fox, Kyoo il Kim, Stephen P Ryan, and Patrick Bajari. The random coefficients logit model is identified. *Journal of Econometrics*, 166(2):204–212, 2012.
- [10] Kenneth E Train. Em algorithms for nonparametric estimation of mixing distributions. *Journal of Choice Modelling*, 1(1):40–69, 2008.
- [11] Lily Hong Zhang, Smitha Milli, Karen Jusko, Jonathan Smith, Brandon Amos, Wassim Bouaziz, Manon Revel, Jack Kussman, Yasha Sheynin, Lisa Titus, Bhaktipriya Radharapu, Jane Yu, Vidya Sarma, Kris Rose, and Maximilian Nickel. Cultivating pluralism in algorithmic monoculture: The community alignment dataset. *arXiv preprint arXiv:2507.09650*, 2025.

A Appendix

A.1 Identifiability

Denote $P(i|\mathcal{S}_i, x)$ to be the probability that i is preferred to all elements in available set $\mathcal{S}_i$, given prompt x . $P(i|\mathcal{S}_i, x, \alpha)$ to be the probability that i is preferred to all elements in $\mathcal{S}_i$, given the distribution of α and prompt x . Here we use random-logit kernel and utility function $U_{\alpha, \theta_1, \theta_2}$ for $P(i|\mathcal{S}_i, x, \alpha)$:

$$P(i|\mathcal{S}_i, x) = \int P(i, |\mathcal{S}_i, x, \alpha) f(\alpha) d\alpha = \int \frac{\exp(U_{\alpha, \theta_1, \theta_2}(i))}{\sum_{j \in \mathcal{S}_i} \exp(U_{\alpha, \theta_1, \theta_2}(j))} f(\alpha) d\alpha. \quad (16)$$

We also assume that we have the known feature map for embedding and **reward functions are single-index**. Let $\Phi = [\phi_1, \phi_2, \dots, \phi_n]^T \in \mathbb{R}^{n \times k}$ denote a known feature matrix for a given prompt x , where $\phi_i \in \mathbb{R}^k$ is the feature vector of response i . Let $\theta_1, \theta_2 \in \mathbb{R}^k$ parameterize the two reward functions R_1 and R_2 . With a single-index form

$$R_\ell(x, i) = \phi_i^\top \theta_\ell, \ell \in \{1, 2\}.$$

Note that here we use the same feature map ϕ for both R_1 and R_2 for simpler notation, but this is not necessary. With this re-parametrization

$$\begin{aligned} U_{\alpha, \theta_1, \theta_2}(x, i) &= \alpha R_1(x, i) + (1 - \alpha) R_2(x, i) \\ &= \alpha \phi_i^\top \theta_1 + (1 - \alpha) \phi_i^\top \theta_2 \\ &= \phi_i^\top (\theta_2 + \alpha(\theta_1 - \theta_2)). \end{aligned} \quad (17)$$

Define $\beta = \theta_2 + \alpha(\theta_1 - \theta_2) \doteq \theta_2 + \alpha d$, our probability model now is

$$P(i, |\mathcal{S}_i, x) = \int \frac{\exp(\phi_i^\top \beta)}{\sum_{j \in \mathcal{S}_i} \exp(\phi_j^\top \beta)} f_\beta(\beta) d\beta. \quad (18)$$

According to theorem 4.1 in paper [9], $f_\beta(\beta)$ is **identifiable** when we have **at least 3 options** in each available set $\mathcal{S}_i$. In other words, we need **best of three** preference data instead of only pairwise comparison data.

A.2 Example

1. **Prompt:** Explain the importance of utilizing social media influences in this market?
2. Preferred by our **user- α model**, which also is the true **winning answer**: Utilizing social media influencers in the handmade jewelry market is important because it allows for targeted and niche marketing. Different influencers cater to different demographics, interests, and styles, enabling you to reach specific segments of the market that align with your brand. For example, if your handmade jewelry line focuses on eco-friendly and sustainable materials, you can partner with influencers who specialize in green living or ethical fashion. This

targeted approach ensures that your marketing efforts are seen by people who are already interested in what you offer, increasing the likelihood of conversion and brand loyalty. Additionally, influencer marketing can be more cost-effective than traditional forms of advertising, providing a higher return on investment for small businesses and independent artisans.

3. Preferred by **global baseline**: Social media influencers are essential in the handmade jewelry market because they offer a form of social proof that traditional advertising cannot match. When an influencer endorses a product, their followers are more likely to trust the recommendation and consider purchasing the item. This is especially true for handmade jewelry, which often involves a higher price point and a more personal connection with the customer. Influencers can showcase your products in a lifestyle setting, demonstrating how the jewelry can be styled and worn, which helps potential customers visualize the product and makes it more desirable. Furthermore, influencers can provide feedback and insights on your products, helping you refine your designs and marketing strategies.
4. Preferred by **local baseline**: The importance of social media influencers in the handmade jewelry market lies in their ability to create an emotional connection with their audience. Influencers often share stories about the brands they work with, highlighting the craftsmanship, sustainability, and ethical practices behind the products. This narrative-driven approach resonates deeply with consumers who are looking for more than just a product – they want to support brands that share their values. By collaborating with influencers who embody these values, you can create a sense of community and social responsibility around your brand, attracting customers who are loyal and passionate about handmade jewelry.