Modeling the Plurality of Human Preferences via Ideal Points

Daiwei Chen [1] **Yi Chen** [1] **Aniket Rege** [2] **Ramya Korlakai Vinayak** [1]

Abstract

Large foundation models require extensive *alignment* to human preferences before deployment. Existing methods utilize the Bradley-Terry-Luce (BTL) model (Bradley & Terry, 1952) often assume a universal preference, neglecting the diversity of individual opinions. We introduce PAL, a framework that models the *plurality* of human preferences using the ideal point model and mixture modeling. PAL captures this *plurality* while learning a common preference latent space, enabling few-shot generalization to new users. With simple multi-layer perceptron, PAL achieves competitive reward model accuracy on Summary (Stiennon et al., 2020) (language), Pick-a-Pic (Kirstain et al., 2024) (image generation), and Persona (Perez et al., 2022) (semi-synthetic) heterogeneous preference datasets, matching state-of-the-art performance with greater efficiency. Lastly, our findings highlight the need for more nuanced data collection to capture the heterogeneity of human preferences.

1. Introduction

Large pre-trained "foundation" models (Bommasani et al., 2021), such as large language models (LLMs) for language generation (Achiam et al., 2023; Anil et al., 2023; Anthropic, 2024; Hoffmann et al., 2022; Rae et al., 2021; Reid et al., 2024; Touvron et al., 2023) and text-to-image (TTI (Luccioni et al., 2023)) models for image generation (Ding et al., 2022; Kang et al., 2023; Ramesh et al., 2022; Rombach et al., 2022; Saharia et al., 2022; Sauer et al., 2023; Yu et al., 2022), are trained on massive amounts of data, including data from the internet. While such models learn useful representations for general language or vision tasks, they are not readily deployable out-of-the-box to be used in the real world. Modern machine learning (ML) systems built on large foundation models go through rigorous fine-tuning or *aligning* towards human preferences to make them amenable to real-world usage. This is usually achieved through supervised fine-tuning (SFT) with direct human input on what the desired outputs should look like for a given context and then followed by alignment with large amounts of human preference feedback usually in the form of pairwise comparison of two outputs to a given input context (Ouyang et al., 2022). This is usually achieved either by (i) fine-tuning the SFT model with explicitly learned reward as done in reinforcement learning with human feedback (RLHF) methods such as proximal policy optimization (PPO) (Schulman et al., 2017) or implicitly with methods such as direct preference optimization (DPO) (Rafailov et al., 2024), or (ii) inference-time policy adaptation (Lu et al., 2023) without fine-tuning the original large policy model.

While aligning ML/AI models to human preferences, it is imperative to ask ourselves *whose preferences are we aligning the ML/AI models to?* (Santurkar et al., 2023) The status quo of the alignment phase is to assume a homogeneous preference shared by all humans and attempt to learn a reward model to learn this preference with the Bradley-Terry-Luce (BTL) model (Bradley & Terry, 1952) of paired preferences. We challenge these notions in an attempt to capture diverse, heterogeneous preferences (Bakker et al., 2022; Durmus et al., 2024; Nadal & Chatterjee, 2019; Wildavsky, 1987). The importance of capturing the plurality of preferences and values among humans has also been highlighted recently by Sorensen et al. However, the methods suggested therein and other recent works that look at learning multiple rewards as a top-down approach where the system designer decides the number and axes that one should care about (Cheng et al., 2023; Choi & Li, 2024; Kovač et al., 2023; Ouyang et al., 2022; Santurkar et al., 2023), e.g., helpfulness vs. harmfulness (Bai et al., 2022b;a; Ganguli et al., 2022; Rame et al., 2024). In reality, human preference is more complex than the designer-specified axes (Bakker et al., 2022), which leads us to propose the following goal.

Goal: Develop a framework for pluralistic alignment that uses diverse human preferences from the ground up.

Our Contributions. Towards this goal, we make the following contributions,

[1]Department of Electrical and Computer Engineering, University of Wisconsin-Madison, Madison, WI, USA [2]Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI, USA. Correspondence to: Daiwei Chen <daiwei.chen@wisc.edu>.

Work presented at TF2M workshop at ICML 2024, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1. **Novel Reformulation:** We reframe the problem of alignment from human preferences by introducing the lens of ideal point model (Coombs, 1950) and metric learning (Kulis, 2013).

2. **New Framework for Pluralistic Alignment:** We propose PAL, a general framework for pluralistic alignment using diverse human preferences from the ground up.

3. **Empirical Validation on Benchmark Datasets**: We evaluate PAL through extensive experiments (Section 3) on both synthetic and real datasets.

2. Framework for Pluralist Alignment (PAL)

In this section, we describe how to view existing approaches that use the BTL model for alignment through the lens of the ideal point model, and then introduce our framework for pluralistic alignment. We illustrate the PAL framework in Figure 1 and Figure 2 (Appendix C). See Appendix B for problem background and notation used below.

2.1. Viewing alignment through the lens of ideal point model and metric learning

The ideal point model assumption that items lie in a vector space is reasonable, especially with foundation models. However, the assumption that a known distance function in this space accurately captures human similarity judgments is a strong one. We reframe alignment (learning a reward function) as learning a transformation of the foundation model's representation space, where a known distance function (e.g. Euclidean distance) better approximates human similarity judgments in this transformed space.

Looking at the current alignment approaches using the BTL model through the lens of the ideal point model, we can reinterpret $1/(1+\exp(r_\theta(\mathbf{x}_r;\mathbf{x}_c)-r_\theta(\mathbf{x}_l;\mathbf{x}_c)))$ as an ideal point model where the difference of rewards is a proxy for the difference of distances[1] and the link function being the Sigmoid or logistic function.

We relax the requirements of 1) Sigmoid link function used by the BTL model and 2) known distance function used by the ideal point model and propose to view alignment from human preferences as learning a reward function that can generalize to the following:

$$\Pr(\mathbf{x}_l \succ \mathbf{x}_r|\mathbf{x}_c) = h(r_\theta(\mathbf{x}_r;\mathbf{x}_c) - r_\theta(\mathbf{x}_l;\mathbf{x}_c)), \quad (1)$$

where h any monotonic link function appropriately normalized to obtain probabilities.

We instantiate the reward function in the following ways:

[1]We note that here reward function is a proxy and not real distance function.

1. With *unknown* but fixed ideal point, *unknown* representation space for *jointly* representing the prompt input $\mathbf{x}_c$ and the corresponding output $\mathbf{x}$ from the foundation model and *Euclidean distance*, we obtain the formulation, $r_\theta(\mathbf{x},\mathbf{x}_c) = ||f(\mathbf{x};\mathbf{x}_c) - f(\mathbf{a})||^2$, where mapping $f:\mathbb{R}^{2D}\to\mathbb{R}^d$ and ideal point $\mathbf{a}\in\mathbb{R}^{2D}$ are *unknown* and learned from pairwise comparison queries. This corresponds to the following pairwise ranking model:

$$\begin{aligned}&\Pr(\mathbf{x}_l \succ \mathbf{x}_r|\mathbf{x}_c)\\ &= h(||f(\mathbf{x}_r;\mathbf{x}_c) - f(\mathbf{a})||_2^2 - ||f(\mathbf{x}_l;\mathbf{x}_c) - f(\mathbf{a})||_2^2)\end{aligned}$$

2. The user ideal point is an *unknown* function of the prompt $\mathbf{x}_c$ and distance is the angle between the ideal point conditioned on the prompt and the *unknown* representation space for the output $\mathbf{x}$ from the foundation model, $r_\theta(\mathbf{x},\mathbf{x}_c) = \langle f(\mathbf{x}), z(\mathbf{x}_c)\rangle$, where the mappings f and z map $\mathbb{R}^D\to\mathbb{R}$ and are *unknown* and are learned from pairwise comparisons. Here we assume that the range spaces of f and z are normalized to use the angle as the distance function. This corresponds to the following pairwise ranking model:

$$\Pr(\mathbf{x}_l \succ \mathbf{x}_r|\mathbf{x}_c) = h(\langle f(\mathbf{x}_r), z(\mathbf{x}_c)\rangle) - \langle f(\mathbf{x}_l), z(\mathbf{x}_c)\rangle)$$

2.2. Modeling diverse preferences

So far we have focused on viewing current alignment methods, which assume a *homogeneous* model i.e. all users' preferences are assumed to arrive from a universal model with disagreements modeled as noise. A natural extension to individualized modeling can be written as follows: For user i, $\Pr_i(\mathbf{x}_l \succ \mathbf{x}_r|\mathbf{x}_c) = h^{(i)}(r_\theta^{(i)}(\mathbf{x}_r;\mathbf{x}_c) - r_\theta^{(i)}(\mathbf{x}_l;\mathbf{x}_c))$,where $h^{(i)}$ is any monotonic link function that can be dependent on the individual and $r^{(i)}(.)$ denotes the reward function for individual i. We note that the learning algorithm does not need to know the link function. One could use these models at a single-user level to learn a personalized model using lots of data from that specific user. However, such models will not generalize to other individuals.

In reality, individuals' preferences exhibit systematic differences and shared aspects within subgroups (e.g., due to demographics). We propose a low-rank mixture model to capture this diversity, representing each user as a convex combination of K prototypical preferences.

Model A: Diverse preference with fixed preference points. Here each user's ideal point is modeled as a convex combination of K prototypical ideal points, $\{\mathbf{p}_1,...,\mathbf{p}_K\}$ with $\mathbf{p}_i\in\mathbb{R}^{2D}$.

$$\begin{aligned}\textbf{Model A:}\quad &\Pr_i(\mathbf{x}_l \succ \mathbf{x}_r|\mathbf{x}_c)\\ &= h(||f(\mathbf{x}_r;\mathbf{x}_c) - f(\mathbf{a}^{(i)})||_2^2\\ &\quad - ||f(\mathbf{x}_l;\mathbf{x}_c) - f(\mathbf{a}^{(i)})||_2^2),\end{aligned} \quad (2)$$

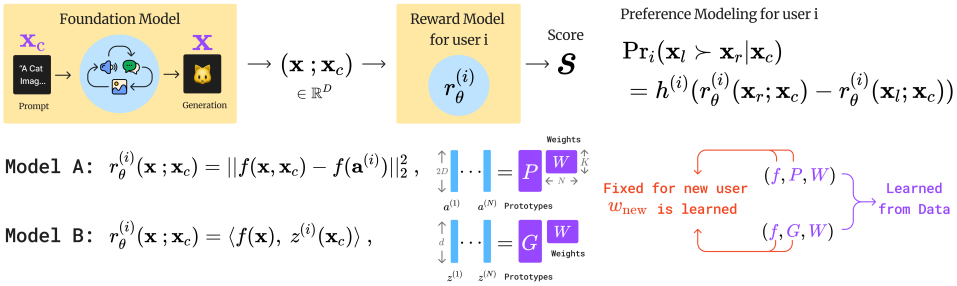


Figure 1: Illustration of PAL framework for learning from diverse preferences (Section 2). For any user i, the probability of preferring $\mathbf{x}_l$ to $\mathbf{x}_r$ for the context $\mathbf{x}_c$ is given by a reward model $r_\theta^{(i)}$ which is modeled with a mixture modeling approach to capture diverse user preferences – each user's preference is modeled as a convex combination of K prototypes. Reward function formulated using PAL framework can be used flexibly, e.g, with fixed preference points (Model A) or with preference points that are functions of the context/prompt $\mathbf{x}_c$ (Model B).

where $\mathbf{a}^{(i)} := \sum_{k=1}^{K} w_k^{(i)} \mathbf{p}_k$ with the weights $w_k^{(i)} \geq 0$ and $\sum_{k=1}^{K} w_k^{(i)} = 1$. Denoting $\mathbf{P} := [\mathbf{p}_1, \cdots, \mathbf{p}_K]$ and $\mathbf{w}^{(i)} := [w_1^{(i)}, \cdots, w_K^{(i)}]^\top$, $\mathbf{a}^{(i)} = \mathbf{P}\mathbf{w}^{(i)}$, where $\mathbf{w}^{(i)}$ lies in K-dimensional simplex denoted by Δ^K.

Model B: Diverse preference with preference points as function of input prompt. Here each user's ideal point is modeled as a convex combination of K prototypical functions that map input prompts to *ideal points*, $\{g_1, ..., g_K\}$.

$$\textbf{Model B:} \quad \Pr_i(\mathbf{x}_l \succ \mathbf{x}_r | \mathbf{x}_c) = h\left(\langle f(\mathbf{x}_r), z^{(i)}(\mathbf{x}_c)\rangle - \langle f(\mathbf{x}_l), z^{(i)}(\mathbf{x}_c)\rangle\right),$$

where $z^{(i)}(\mathbf{x}_c) = \sum_{k=1}^{K} w_k^{(i)} g_k(\mathbf{x}_c) = \mathbf{G}(\mathbf{x}_c)\mathbf{w}^{(i)}$ with $\mathbf{G}(\mathbf{x}_c) := [g_1(\mathbf{x}_c), \cdots, g_K(\mathbf{x}_c)]$ and $\mathbf{w}^{(i)} \in \Delta^K$. We drop the superscript i on h for simplicity, however, we note that the link function need not be the same for all users, and furthermore, our learning algorithm described in Section 2.3 does not need to know the link function(s). We illustrate the PAL framework in Figure 1 and Figure 2 (Appendix C).

2.3. Learning PAL models from Diverse Preferences

Given a dataset of answers to pairwise comparison queries, $\left\{\{(\mathbf{x}_l, \mathbf{x}_r; \mathbf{x}_c)_j^{(i)}\}_{j=1}^{m_i}\right\}_{i=1}^{N}$, where m_i denote the number of pairs answered by user i, the **goal** of the learning algorithm in the PAL framework is to learn the mappings and prototypes shared across the population, and for each user i the weights $\mathbf{w}^{(i)} := [w_1^{(i)}, ..., w_K^{(i)}]$ with $w_k^{(i)} \geq 0$ and $\sum_{k=1}^{K} w_k^{(i)} = 1$. For model A, mapping f and the prototypes $\{\mathbf{p}_k\}_{k=1}^{K}$ are shared, while for model B, the mapping f and the prototype mappings $\{g_k\}_{k=1}^{K}$ shared. Without loss of generality, we have assumed that $\mathbf{x}_l$ is preferred over $\mathbf{x}_r$. Thus, this learning problem can be viewed as a supervised learning setting with binary labels.

Generalization over *seen* users versus *unseen* users: When learning a reward function from diverse preferences, there are two types of generalization to consider. (1) Generalization for *unseen pairs for seen users*, i.e. predicting well for new pairs for the people for whom the weights have already been learned from the training data. We call this *seen accuracy*. (2) Generalization for *unseen users*, i.e. predicting well for people whose data was not part of the training data at all. For such new users, some part of their new data will be used to localize their weights, with the shared mappings and prototypes fixed. We call this *unseen accuracy*. We also note that we can use the weighted combination of learned prototypes, i.e. an average of all the seen users, as the *zero-shot* ideal point for new users. However, we emphasize that it is important for reward functions to generalize to *unseen* users and our framework provides a natural way to localize a new user.

Algorithm. Given the following **input:** Dataset $\mathcal{D} = \left\{\{(\mathbf{x}_l, \mathbf{x}_r; \mathbf{x}_c)_{j_i}^{(i)}\}_{j_i=1}^{m_i}\right\}_{i=1}^{N}$, loss function ℓ and model class for f_θ, the learning algorithm for **Model A** starts by randomly initializing the prototypes $\mathbf{P} = [\mathbf{p}_1, ..., \mathbf{p}_K]$, $\mathbf{p}_k \in \mathbb{R}^d$, user weights $\mathbf{W} = [\mathbf{w}^{(1)}, ..., \mathbf{w}^{(N)}]$, where $\mathbf{w}^{(i)} \in \Delta^K$. Then, in each iteration until convergence criteria, the following steps are repeated (see Figure 1):

- **Sample** a random mini-batch $\left\{(\mathbf{x}_l, \mathbf{x}_r; \mathbf{x}_c)_j^{(i)}\right\}$ of comparison data from $\mathcal{D}$.
- **Compute user ideal points:** $\mathbf{a}^{(i)} = \mathbf{P} \cdot \mathbf{w}^{(1)}$
- **Compute distances**: $d_{l,j}^{(i)} = ||f_\theta(\mathbf{x}_l; \mathbf{x}_c) - f_\theta(\mathbf{a}^{(i)})||_2^2$, $d_{r,j}^{(i)} = ||f_\theta(\mathbf{x}_r; \mathbf{x}_c) - f_\theta(\mathbf{a}^{(i)})||_2^2$.
- **Loss for each comparison** j **for user** i: $\ell_j^{(i)}(\mathbf{x}_l, \mathbf{x}_r; \mathbf{x}_c) = \ell(d_{r,j}^{(i)} - d_{l,j}^{(i)})$.
- **Update Step:** $\mathbf{argmax}_{\theta, \mathbf{P}, \{\mathbf{w}^{(i)}\}_{i=1}^{N}} \sum_{i,j} \ell_j^{(i)}(\mathbf{x}_l, \mathbf{x}_r; \mathbf{x}_c)$.

The above steps describe updating the learning algorithm for model A. For model B, the steps are similar except that

prototypes now are the functions (g_k) and the distance is the angle ($\langle\cdot\rangle$). See Appendix F for pseudocode details.

Table 1: PAL test accuracy can match SoTA Pickscore (Kirstain et al., 2024) on the Pick-a-Pic-v1 test set with a fraction of the compute.

Model	Test Accuracy(%) Pick-a-Pic v1 test
CLIP-H14	59.23
PickScore	**71.85**
model A on CLIP-H	69.29 ± 0.66
model B on CLIP-H	$\mathbf{71.13 \pm 0.31}$

Table 2: Test Accuracy of PAL compared to CLIP-H and Pickscore baselines on Pick-a-Pic v2. Entries with asterisk* have inflated accuracies due to V2 test set overlap with V1 train.

Model	Train set	Test Acc. Pick-a-Pic v2 (%)	
		No-leakage	Leakage
CLIP-H14	-	62.57	58.59
PickScore	pickapic v1	68.04	74.16^*
model B (CLIP-H)	**pickapic v1**	$\mathbf{70.02 \pm 0.39}$	$79.32 \pm 1.68^*$
model B (CLIP-H)	**pickapic v2**	$\mathbf{70.51 \pm 0.22}$	$\mathbf{68.67 \pm 0.51}$
model B (PickScore)	pickapic v2	70.16 ± 0.19	$74.79 \pm 0.13^*$

3. Experiments

We conduct extensive experiments on both simulated (Appendix E.1) and real preference datasets (Appendix E.2) for both text and image generation tasks to demonstrate that our proposed PAL (Pluralistic ALignment) framework can: (1) effectively capture the diversity of user preferences, thereby outperforming existing homogeneous reward models; (2) efficiently achieve performance comparable to the existing SoTA reward models with far fewer parameters and compute costs (See Appendix G); and (3) versatile to applied to different domains. For experiments on real preference datasets, a simple two-layer MLP PAL reward model can achieve or exceed the performance of the existing status quo reward models, which often contain billions of parameters. Due to the page limitation, we report our results on the Pick-a-Pic dataset and defer the rest to Appendix E.

3.1. Pick-a-Pic Dataset

We conducted experiments on the Pick-a-Pic dataset (Kirstain et al., 2024) and show two benefits of our proposed ideal point model compared with existing reward models i.e. learning 1) diverse user preferences 1) a competitive reward model with only 2-layer MLP networks. Recent works on the existing reward models usually require fine-tuning foundation models with billions of parameters, while PAL achieves comparable performance without a fine-tuning stage, saving significant compute costs.

Dataset and Experiment Setup. There are two versions of Pick-a-Pic datasets, v1 and v2 (which extends v1). We trained model B on both datasets, using CLIP-H14 or PickScore latent embeddings as input. Due to sample overlapping between the v1 training set and the v2 test set, we split the v2 test into "no-leakage" and "leakage" subsets to fairly compare our model with the SoTA PickScore reward model, which is trained on v1. We adopt the same hyperparameters used in *Pick-a-Filter* experiments (Appendices D and E.2.2), avoiding extensive hyperparameter tuning. Our model B is trained on CLIP-H14 or PickScore latent embeddings from either the v1 or v2 datasets, over 10 epochs.

Results. Table 1 demonstrates that the performance of our model B aligns with SoTA PickScore on the Pick-a-Pic dataset. On the v2 no-leakage test set, our model B outperforms PickScore by **2%** (Table 2). Additionally, the performance of model B using PickScore latent embeddings is inferior to that of model B using CLIP-H14 embeddings. This highlights the effectiveness of our proposed ideal point model framework: it can match or exceed the SoTA reward model using a simple two-layer MLP network, whereas PickScore requires fine-tuning the entire CLIP-H14 model ($\sim 1B$ parameters) with 8×A100 GPUs.

Remark. Since the data collection process for existing datasets involves the usage of strict rubrics (Kirstain et al., 2024; Stiennon et al., 2020; Wu et al., 2023), labeler performance monitoring (Xu et al., 2024) and disproportionate amount of data from small fraction of users, these datasets may not be heterogeneous. Using PAL with $K = 1$, we surpass existing SoTA (Table 2). These results motivate the need to collect datasets that contain unmoderated, diverse opinions.

4. Conclusions, Limitations and Future Work

We proposed a novel reformulation of the problem of alignment with human preferences via the PAL framework for *pluralistic alignment* with diverse preferences from the ground up (Section 2). PAL leverages shared structures across the user population while learning to personalize to individuals using a mixture modeling approach. We demonstrate the PAL framework is agnostic to modality, showing flexible adaptivity to heterogeneous preferences on text data and image data (Section 3). Our work aids in building much-needed foundations toward plurality for alignment of ML/AI models. Our experiments also highlight the limitations of many real human preference datasets that are collected with rubrics that make the dataset homogeneous, thus calling for a more nuanced approach to data collection in the future.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

Anthropic, A. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 2024.

Azar, M. G., Guo, Z. D., Piot, B., Munos, R., Rowland, M., Valko, M., and Calandriello, D. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pp. 4447–4455. PMLR, 2024.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Bakker, M., Chadwick, M., Sheahan, H., Tessler, M., Campbell-Gillingham, L., Balaguer, J., McAleese, N., Glaese, A., Aslanides, J., Botvinick, M., et al. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189, 2022.

Bellet, A. and Habrard, A. Robustness and generalization for metric learning. *Neurocomputing*, 151:259–267, 2015.

Bellet, A., Habrard, A., and Sebban, M. Metric learning. *Synthesis lectures on artificial intelligence and machine learning*, 9(1):1–151, 2015.

Bellet, A., Habrard, A., and Sebban, M. *Metric learning*. Springer Nature, 2022.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

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

Braverman, M. and Mossel, E. Noisy sorting without resampling. *arXiv preprint arXiv:0707.1051*, 2007.

Canal, G., Mason, B., Vinayak, R. K., and Nowak, R. One for all: Simultaneous metric and preference learning over multiple users. *arXiv preprint arXiv:2207.03609*, 2022.

Cheng, M., Durmus, E., and Jurafsky, D. Marked personas: Using natural language prompts to measure stereotypes in language models. *arXiv preprint arXiv:2305.18189*, 2023.

Choi, H. K. and Li, Y. Beyond helpfulness and harmlessness: Eliciting diverse behaviors from large language models with persona in-context learning. *arXiv preprint arXiv:2405.02501*, 2024.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Coombs, C. H. Psychological scaling without a unit of measurement. *Psychological review*, 57(3):145, 1950.

Ding, C. Evaluating change in behavioral preferences: Multidimensional scaling single-ideal point model. *Measurement and Evaluation in Counseling and Development*, 49 (1):77–88, 2016.

Ding, M., Zheng, W., Hong, W., and Tang, J. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35:16890–16902, 2022.

Dong, H., Xiong, W., Goyal, D., Zhang, Y., Chow, W., Pan, R., Diao, S., Zhang, J., Shum, K., and Zhang, T. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023.

Dulac-Arnold, G., Levine, N., Mankowitz, D. J., Li, J., Paduraru, C., Gowal, S., and Hester, T. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning*, 110(9):2419–2468, 2021.

Durmus, E., Nguyen, K., Liao, T. I., Schiefer, N., Askell, A., Bakhtin, A., Chen, C., Hatfield-Dodds, Z., Hernandez, D., Joseph, N., Lovitt, L., McCandlish, S., Sikder, O., Tamkin, A., Thamkul, J., Kaplan, J., Clark, J., and Ganguli, D. Towards measuring the representation of subjective global opinions in language models, 2024.

Eriksson, B. Learning to top-k search using pairwise comparisons. In *Artificial Intelligence and Statistics*, pp. 265–273. PMLR, 2013.

Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., and Kiela, D. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

Fürnkranz, J. and Hüllermeier, E. Preference learning and ranking by pairwise comparison. In *Preference learning*, pp. 65–82. Springer, 2010.

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Huber, J. Ideal point models of preference. *ACR North American Advances*, 1976.

Hunter, D. R. Mm algorithms for generalized bradley-terry models. *The annals of statistics*, 32(1):384–406, 2004.

Irvine, R., Boubert, D., Raina, V., Liusie, A., Zhu, Z., Mudupalli, V., Korshuk, A., Liu, Z., Cremer, F., Assassi, V., et al. Rewarding chatbots for real-world engagement with millions of users. *arXiv preprint arXiv:2303.06135*, 2023.

Jamieson, K. G. and Nowak, R. Active ranking using pairwise comparisons. *Advances in neural information processing systems*, 24, 2011.

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

Kang, M., Zhu, J.-Y., Zhang, R., Park, J., Shechtman, E., Paris, S., and Park, T. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10124–10134, 2023.

Kenyon-Mathieu, C. and Schudy, W. How to rank with few errors. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pp. 95–103, 2007.

Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., and Levy, O. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.

Kleindessner, M. and Luxburg, U. Uniqueness of ordinal embedding. In *Conference on Learning Theory*, pp. 40–67. PMLR, 2014.

Kovač, G., Sawayama, M., Portelas, R., Colas, C., Dominey, P. F., and Oudeyer, P.-Y. Large language models as superpositions of cultural perspectives. *arXiv preprint arXiv:2307.07870*, 2023.

Kulis, B. Metric learning: A survey. *Foundations and Trends® in Machine Learning*, 5(4):287–364, 2013.

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., and Legg, S. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

Li, X., Lipton, Z. C., and Leqi, L. Personalized language modeling from personalized human feedback. *arXiv preprint arXiv:2402.05133*, 2024.

Lu, X., Brahman, F., West, P., Jung, J., Chandu, K., Ravichander, A., Ammanabrolu, P., Jiang, L., Ramnath, S., Dziri, N., et al. Inference-time policy adapters (ipa): Tailoring extreme-scale lms without fine-tuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6863–6883, 2023.

Luccioni, A. S., Akiki, C., Mitchell, M., and Jernite, Y. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*, 2023.

Luce, R. D. *Individual choice behavior: A theoretical analysis*. New York: Wiley, 1959.

Mason, B., Jain, L., and Nowak, R. Learning low-dimensional metrics. *Advances in neural information processing systems*, 30, 2017.

Nadal, M. and Chatterjee, A. Neuroaesthetics and art's diversity and universality. *Wiley Interdisciplinary Reviews: Cognitive Science*, 10(3):e1487, 2019.

Negahban, S., Oh, S., and Shah, D. Iterative ranking from pair-wise comparisons. *Advances in neural information processing systems*, 25, 2012.

Nelder, J. A. and Wedderburn, R. W. M. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972. URL http://www.jstor.org/stable/2344614.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Palmer, S. E. and Schloss, K. B. Human preference for individual colors. In *Human Vision and Electronic Imaging XV*, volume 7527, pp. 353–364. SPIE, 2010.

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., Amodei, D., Drain, D., Li, D., Tran-Johnson, E., Khundadze, G., Kernion, J., Landis, J., Kerr, J., Mueller, J., Hyun, J., Landau, J., Ndousse, K., Goldberg, L., Lovitt, L., Lucas, M., Sellitto, M., Zhang, M., Kingsland, N., Elhage, N., Joseph, N., Mercado, N., DasSarma, N., Rausch, O., Larson, R., McCandlish, S., Johnston, S., Kravec, S., El Showk, S., Lanham, T., Telleen-Lawton, T., Brown, T., Henighan, T., Hume, T., Bai, Y., Hatfield-Dodds, Z., Clark, J., Bowman, S. R., Askell, A., Grosse, R., Hernandez, D., Ganguli, D., Hubinger, E., Schiefer, N., and Kaplan, J. Discovering language model behaviors with model-written evaluations, 2022. URL https://arxiv.org/abs/2212.09251.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Rajkumar, A. and Agarwal, S. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *International conference on machine learning*, pp. 118–126. PMLR, 2014.

Rame, A., Couairon, G., Dancette, C., Gaya, J.-B., Shukor, M., Soulier, L., and Cord, M. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems*, 36, 2024.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T., Alayrac, J.-b., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494, 2022.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose opinions do language models reflect? In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29971–30004. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/santurkar23a.html.

Sauer, A., Karras, T., Laine, S., Geiger, A., and Aila, T. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. *arXiv preprint arXiv:2301.09515*, 2023.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Schultz, M. and Joachims, T. Learning a distance metric from relative comparisons. *Advances in neural information processing systems*, 16, 2003.

Shah, N., Balakrishnan, S., Guntuboyina, A., and Wainwright, M. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. In *International Conference on Machine Learning*, pp. 11–20. PMLR, 2016.

Shah, N. B. and Wainwright, M. J. Simple, robust and optimal ranking from pairwise comparisons. *The Journal of Machine Learning Research*, 18(1):7246–7283, 2017.

Shepard, R. N. The analysis of proximities: multidimensional scaling with an unknown distance function. i. *Psychometrika*, 27(2):125–140, 1962a.

Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. ii. *Psychometrika*, 27(3):219–246, 1962b.

Shepard, R. N. Metric structures in ordinal data. *Journal of Mathematical Psychology*, 3(2):287–315, 1966.

Singla, A., Tschiatschek, S., and Krause, A. Actively learning hemimetrics with applications to eliciting user preferences. In *International Conference on Machine Learning*, pp. 412–420. PMLR, 2016.

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., et al. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021, 2020.

Tamuz, O., Liu, C., Belongie, S., Shamir, O., and Kalai, A. T. Adaptively learning the crowd kernel. *arXiv preprint arXiv:1105.1033*, 2011.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., von Werra, L., Fourrier, C., Habib, N., et al. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*, 2023.

Wang, B., Zheng, R., Chen, L., Liu, Y., Dou, S., Huang, C., Shen, W., Jin, S., Zhou, E., Shi, C., et al. Secrets of rlhf in large language models part ii: Reward modeling. *arXiv preprint arXiv:2401.06080*, 2024a.

Wang, H., Lin, Y., Xiong, W., Yang, R., Diao, S., Qiu, S., Zhao, H., and Zhang, T. Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards. *arXiv preprint arXiv:2402.18571*, 2024b.

Wang, Z., So, G., and Vinayak, R. K. Metric learning from limited pairwise preference comparisons. In *UAI*, 2024c.

Wildavsky, A. Choosing preferences by constructing institutions: A cultural theory of preference formation. *American political science review*, 81(1):3–21, 1987.

Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., and Li, H. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Wu, Z., Hu, Y., Shi, W., Dziri, N., Suhr, A., Ammanabrolu, P., Smith, N. A., Ostendorf, M., and Hajishirzi, H. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36, 2024.

Xu, A. and Davenport, M. Simultaneous preference and metric learning from paired comparisons. *Advances in Neural Information Processing Systems*, 33:454–465, 2020.

Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., and Dong, Y. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.

Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B. K., et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

Yuan, Z., Yuan, H., Tan, C., Wang, W., Huang, S., and Huang, F. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*, 2023.

A. Related Works

Alignment Status Quo. Popular existing foundation models (Achiam et al., 2023; Anthropic, 2024; Ouyang et al., 2022; Touvron et al., 2023) typically use RLHF (Christiano et al., 2017; Stiennon et al., 2020) to align models after pretraining. Recent foundation models such as Zephyr (Tunstall et al., 2023) and the Archangel suite[2] have shifted to directly optimizing on human preferences (Azar et al., 2024; Ethayarajh et al., 2024; Rafailov et al., 2024) to avoid the nuances of RL optimization (Dulac-Arnold et al., 2021). There has also been significant recent work in collecting large human preference datasets for reward model training in the text-to-image (typically diffusion model (Rombach et al., 2022)) space (Kirstain et al., 2024; Wu et al., 2023; Xu et al., 2024).

Reward Modeling. These existing alignment frameworks generally assume that all humans share a single unified preference (e.g. LLM "helpfulness" or "harmlessness" (Bai et al., 2022a)) and ascribe to the Bradley-Terry (Bradley & Terry, 1952) model of pairwise preferences. Consensus-based methods (Bakker et al., 2022) aim to find agreement among labelers for specific goals like harmlessness (Bai et al., 2022b; Ganguli et al., 2022), helpfulness (Bai et al., 2022a), or engagement (Irvine et al., 2023). By design, these methods inherently prioritize the universal preference (and biases) induced by the labelers (Cheng et al., 2023; Kovač et al., 2023; Santurkar et al., 2023). In reality, humans have diverse, heterogeneous preferences (Nadal & Chatterjee, 2019; Sorensen et al., 2024; Wildavsky, 1987) that depend on individual contexts, and may even share a group structure (Bakker et al., 2022). Rewarded soups (Rame et al., 2024) make a case to capture diversity through post-hoc weight-space interpolation over a mixture of experts that learn diverse rewards. However, these rewards are learned by pre-defining what aspects are important which is done by the system designer. Separate datasets are collected to elicit human preferences on these axes as to how much people care of them. DPA (Wang et al., 2024b) models rewards as directions instead of scalars, and trains a multi-objective reward model for RLHF. Wu et al. propose fine-grained multi-objective rewards to provide more focused signal for RLHF. Recently, Li et al. propose personalized reward modeling by learning a general user embedding and treating each individual as a perturbation to the embedding. As this preference formulation is still homogeneous, they can only generalize to unseen users using the fixed general user embedding.

Recent survey works provide excellent summaries of literature for alignment (Ji et al., 2023) and reward modeling (Wang et al., 2024a).

Human Preference Datasets. The preference universality assumption also extends into the data annotation/labeling processing, where labelers are given a rubric to select preferences (e.g. to rank an image pair considering image aesthetics and image-prompt alignment (Kirstain et al., 2024)). Due to this rubric, the current largest scale text-to-image generation preference datasets (Kirstain et al., 2024; Wu et al., 2023; Xu et al., 2024) show limited diversity among labelers. In the Pick-a-Pic (Kirstain et al., 2024) train set, there are only 701 disagreements among the 12487 image pairs labeled by different users (94.38% agreement), and there are zero disagreements in validation (1261 pairs) and test (1453 pairs) sets. HPS (Wu et al., 2023) found that labeler agreement over diffusion model generations was higher for models of similar quality or size, though this diversity comes with the caveat of the labelers being provided a rubric to provide their preferences. Imagereward (Xu et al., 2024) use researcher agreement as a *criteria* to hire labelers. In the LLM domain, the popular Summarize from Feedback dataset (Stiennon et al., 2020) is also collected with rigid rubric, with labeler performance measured via agreement to the preferred answer of the authors. During the data collection period, only labelers with satisfactory agreement were retained, which led to a small number of users, all in agreement with the authors' rubric, being responsible for a majority of labeled comparisons. Status quo preference datasets used to align foundation models thus suffer from a lack of diversity due to the nature of their data collection.

Preference learning. There is rich literature on preference learning and ranking in various domains ranging from psychology, marketing, recommendation systems, quantifying social science surveys to crowdsourced democracy, voting theory and social choice theory. We provide a few relevant works here and direct reader to surveys such as (Fürnkranz & Hüllermeier, 2010). Ranking based models, e.g., BTL-model (Bradley & Terry, 1952; Luce, 1959), stochastic transitivity models (Shah et al., 2016) focus on finding ranking of m items or finding top-k items by pairwise comparisons (Hunter, 2004; Kenyon-Mathieu & Schudy, 2007; Braverman & Mossel, 2007; Negahban et al., 2012; Eriksson, 2013; Rajkumar & Agarwal, 2014; Shah & Wainwright, 2017). Ranking m items in these settings requires $\mathcal{O}(m \log m)$ queries. There is also rich literature that stems from ideal point model proposed by Coombs (Coombs, 1950; Huber, 1976; Jamieson & Nowak, 2011; Ding, 2016; Singla et al., 2016; Xu & Davenport, 2020; Canal et al., 2022). Under the ideal point based models, the query complexity for ranking m items reduces to $\mathcal{O}(d \log m)$, where d is the dimension of the domain of representations which is usually much

[2] https://github.com/ContextualAI/HALOs

smaller than the number of items being ranked (Jamieson & Nowak, 2011). This is due to the fact that once the preference point is learned, it can then be used to predict rankings of new items without needing more comparisons.

Metric learning has been studied quite extensively and we direct the reader to surveys (Kulis, 2013) and books (Bellet et al., 2022). In particular, metric learning based on triplet querying has also been quite extesively studied (Shepard, 1962a;b; 1966; Schultz & Joachims, 2003; Kulis, 2013; Tamuz et al., 2011; Kleindessner & Luxburg, 2014; Bellet et al., 2015; Bellet & Habrard, 2015; Mason et al., 2017) which aims to learn the underlying unknown metric under the assumption that the people base their judgement for a triple query with concepts $\mathbf{x}_a, \mathbf{x}_b, \mathbf{x}_c \in \mathcal{D}$ on the relative similarities based on the distances between these concepts under the unknown metric.

Simultaneous metric and preference learning. More recently a few works have considered the problem of unknown metric in preference learning and proposed methods (Xu & Davenport, 2020; Canal et al., 2022; Wang et al., 2024c) and provided sample complexity analysis (Canal et al., 2022; Wang et al., 2024c) for simultaneously learning an unknown Mahalanobis metric and unknown user preference(s). Learning the unknown Mahalanobis metric can be viewed as learning linear layer on top of the embeddings from a foundation model. From our reframing of alignment, these works can be looked as model A with linear function for f and individual user preferences instead of having any structure over them.

B. Notations and Background

We begin with a brief discussion of the BTL model and how it is currently used in reward learning from pairwise preference comparisons, followed by motivating the ideal point model.

Bradley-Terry-Luce (BTL) model (Bradley & Terry, 1952) is a parametric model for ranking. Given m items or alternatives, the assumption is that there is a universal ranking: $\sigma(1) \succ \sigma(2) \succ \cdots \succ \sigma(m)$ which are a reflection of the *unknown* true scores or weights associated with each of these items $s^\star_{\sigma(1)} > s^\star_{\sigma(2)} > \cdots s^\star_{\sigma(m)}$, where $\sigma(.)$ denotes permutation and the scores $s^\star$ are positive real numbers. Then, the probability that "*i beats j*" when comparing them, denoted by $i \succ j$ is given by,

$$\Pr(i \succ j) = \frac{s_i^\star}{s_i^\star + s_j^\star} = \frac{\exp(r_i)}{\exp(r_i) + \exp(r_j)}, \tag{3}$$

where the variables r re-parameterize $s > 0$.

Notation: We set up some notation for further discussion. Let D denote the dimension of the representation space of the foundation models. Let $\mathbf{x}_c \in \mathbb{R}^D$ denote the representation of the prompt or the context. Let $\mathbf{x}_l \in \mathbb{R}^D$ and $\mathbf{x}_r \in \mathbb{R}^D$ denote the embeddings of two items where the subscripts denote *left* and *right* respectively.

In the literature on alignment with human feedback, the scores re-parametrized with *reward*, denoted here by r, are modeled using a neural network denoted by r_θ. More concretely, given a context or prompt $\mathbf{x}_c$, the probability that output $\mathbf{x}_l$ is preferred to output $\mathbf{x}_r$ under the BTL model is given by,

$$\Pr(\mathbf{x}_l \succ \mathbf{x}_r | \mathbf{x}_c) = \frac{\exp(r_\theta(\mathbf{x}_l; \mathbf{x}_c))}{\exp(r_\theta(\mathbf{x}_l; \mathbf{x}_c)) + \exp(r_\theta(\mathbf{x}_r; \mathbf{x}_c))} \tag{4}$$

$$= \frac{1}{1 + \exp(r_\theta(\mathbf{x}_r; \mathbf{x}_c) - r_\theta(\mathbf{x}_l; \mathbf{x}_c))}. \tag{5}$$

The goal then is to *learn* this *reward function* r_θ that maps the output $\mathbf{x} \in \mathbb{R}^D$ for a given context $\mathbf{x}_c \in \mathbb{R}^D$, denoted by $(\mathbf{x}; \mathbf{x}_c)$, to a real-valued *reward score* to approximate human preference. This learning of r_θ is done using lots of pairwise comparison data obtained by querying humans. Such a learned reward function can be used to *align* the model (Christiano et al., 2017; Leike et al., 2018; Ouyang et al., 2022), score the generations during inference time to output more aligned answers () and to rank the generations of multiple models (Dong et al., 2023; Yuan et al., 2023). Recent work from Rafailov et al. bypasses the status quo two-stage reward learning + RL pipeline and directly finetune on pairwise preferences, but still implicitly assumes the BTL model for ranking.

While most alignment literature focuses on the BTL modeling approach, we want to draw attention to the *ideal point model* (Coombs, 1950) for preference learning.

Ideal point model was proposed by Coombs for human preference modeling in the psychology literature. The key idea behind this model is to exploit the geometry of the problem, assuming there exists a meaningful representation space for the items/alternates being compared. Let $\mathcal{X} \in \mathbb{R}^D$ denote the domain of feature space of the concepts (items, objects, images,

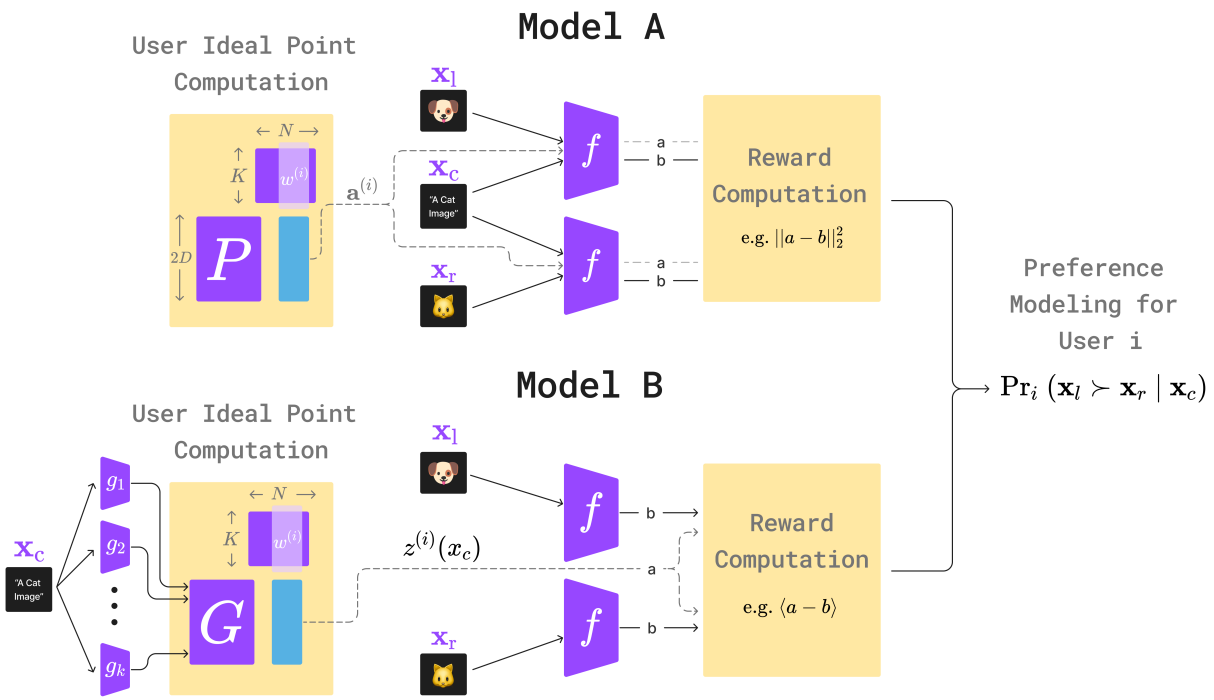


Figure 2: Illustration of PAL framework for learning from diverse preferences (Section 2). For any user i, the probability of preferring $\mathbf{x}_l$ to $\mathbf{x}_r$ for the context $\mathbf{x}_c$ is computed by a reward model $r_\theta^{(i)}$ which uses a mixture modeling approach to assign a scalar reward to a sample (e.g. $\mathbf{x}_l$ or $\mathbf{x}_r$) given context ($\mathbf{x}_c$). In PAL-A, each user i's preference $a^{(i)}$ is modeled as a convex combination of K prototypical preferences, i.e. $a^{(i)} = Pw^{(i)}$. In PAL-B, each user i's preference $z^{(i)}(\mathbf{x}_c)$ is modeled as a convex combination of K prototypical functions $g_1 \cdots g_K$, i.e. $z^{(i)}(\mathbf{x}_c) =$. Reward function formulated using PAL framework can be used flexibly, e.g., with fixed preference points (Model A), with preference points that are functions of the context/prompt $\mathbf{x}_c$ (Model B).

choices, etc.) with a distance associated with it. Preference learning based on ideal point model (Canal et al., 2022; Coombs, 1950; Ding, 2016; Huber, 1976; Jamieson & Nowak, 2011; Singla et al., 2016; Xu & Davenport, 2020) assumes that there is an *unknown* ideal preference point $\mathbf{a} \in \mathcal{X}$ that represents the reference point people use for their preference judgments based on distances. So, when asked "Do you prefer i or j?", they respond with i as their preference if $\text{dist}(\mathbf{x}_i, \mathbf{a}) < \text{dist}(\mathbf{x}_j, \mathbf{a})$ and vice versa, where $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$ are the feature representations of i and j respectively. That is, items that are closer to the user's ideal preference points are preferred by the user over those that are farther away. The goal of preference learning is to use the responses for pairwise comparison queries from people and learn the preference point $\mathbf{a}$. Once we learn $\mathbf{a}$, we can predict the choices people make between new unseen pairs. More formally, in general, the probability that i beats j in preference for user a is given by,

$$\Pr(i \succ j) \propto h(\text{dist}^2(\mathbf{x}_j, \mathbf{a}) - \text{dist}^2(\mathbf{x}_i, \mathbf{a})), \tag{6}$$

where h is a link function (Nelder & Wedderburn, 1972) which can be any monotonic function. Essentially, the idea here is that the larger the difference in distance between the alternates, the easier it is to decide and hence the answer is less noisy. In contrast, if the difference of distance is zero or closer to zero, that means that the alternates seem to be equally good to the user and therefore the probability of $i \succ j$ will be close to random.

C. Model Design

We illustrate the modeling mechanism of PAL (Section 2.2) in slightly more detail in Figure 2.

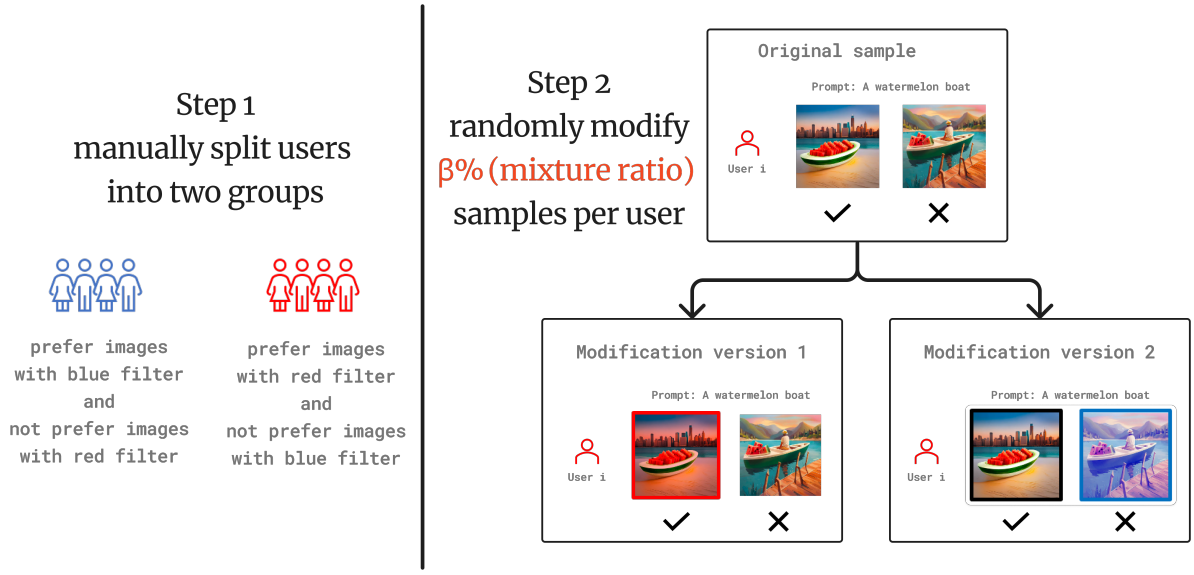


Figure 3: The construction diagram for the semi-synthetic Pick-a-Pic dataset. It involves randomly selecting approximately 100,000 samples from the Pick-a-Pic dataset and dividing the user IDs into two disjoint groups. We assume one group prefers images with "cold" (blue) filters and the other with "warm" (red) filters. To incorporate diverse color filter preferences, we randomly select $\beta\%$ of samples per user on which to apply filters.

D. Dataset Design

Pick-a-Filter : due to the high level of "agreement" among labelers over image preferences on Pick-a-Pic V1 (Kirstain et al., 2024), we construct a semi-synthetic dataset by applying filters to a subset of Pick-a-Pic V1, which we call the Pick-a-Filter dataset. To construct the dataset, we consider only samples that have no ties, i.e. the labeler decides that one image is decisively preferable to the other, given the text prompt. As Pick-a-Pic provides unique and anonymous user IDs for all preference pairs, we consider a subset of users who provide samples in **both** the train and test sets (468 / 4223 users). We further only consider users who provide more than 50 labels (234 / 468 users) and sort the users by number of samples provided. We split these users into equal groups of 117 each, and we assume without loss of generality that the first group of users (G1) prefers "cold" tones (blue filter) and the second group (G2) prefers "warm" tones (red filter). Lastly, we arbitrarily consider the first 50 users (who provide the most number of samples) as "seen" users, i.e. users that provide samples in both the train and test sets of Pick-a-Filter. We add this seen vs. unseen distinction to evaluate how well PAL can adapt to unseen (i.e. new) users after training. Currently, our experiments on Pick-a-Filter (Section E.2.2) train on V1-train-seen (116031 samples) and evaluate on V1-test-seen (3693 samples). We show the number of samples in each of these splits in Table 3. After constructing splits, we apply the following filtering logic:

1. Apply "winning" and "losing" filters to appropriate images depending on label. For G1 the winning filter is blue, and for G2 the winning filter is red.
2. Randomly shortlist $\beta\%$ of samples to add filters. The remaining $(1-\beta)\%$ of samples will remain unaltered (default images from Pick-a-Pic v1).
3. Randomly select 50% of above-shortlisted samples to apply a filter to only the winning image, and the remaining 50% to apply a filter to only losing image

We add these sources of randomness to make learning preferences on Pick-a-Filter less prone to hacking (e.g. the model could trivially learn to predict an image with a filter as the preferred image).

Table 3: Number of samples in each split of the newly constructed Pick-a-Filter dataset.

Category		Train	Val	Test
Group 1	Seen	58831	628	1597
	Unseen	9527	79	1886
	Total	68358	707	3483
Group 2	Seen	57200	404	2096
	Unseen	9402	52	1812
	Total	66602	456	3908

E. Experiments

E.1. Numerical Simulations

Setup. We simulate a simple preference dataset with the normal distribution (we use a setting similar to (Canal et al., 2022)) and true $f^* : \mathbb{R}^d \to \mathbb{R}^d$ is linear and the weight $\mathbf{W} \sim \mathcal{N}(0, I)$. Let $\mathbf{x}_i \sim \mathcal{N}(0, (1/d)I)$ denote the i_{th} item. Assume K^* user prototypes $\{\mathbf{p}_i\}_{i=1}^{K^*}$, where $\mathbf{p}_i \sim \mathcal{N}(0, (1/d)I)$ with the minimum distance constraint $\|\mathbf{p}_i - \mathbf{p}_j\| \geq \delta,\ \forall i, j \in [K^*], i \neq j$. We consider two settings: 1) a **mixture** setting, where we assume each user is located in the convex hull of K prototypes; 2) a simpler **partition** setting, where we assume N users are evenly sampled from K prototypes, with $\mathbf{a}_i \in \{\mathbf{p}_k\}_{k=1}^{K}$. Each sample is generated as follows: we randomly draw two items $\{\mathbf{x}_l, \mathbf{x}_r\}$ and one user $\mathbf{a}_i$, and label the user's preference as $\text{sign}(\|f^*(\mathbf{x}_l) - f^*(\mathbf{a}_i)\|_2 - \|f^*(\mathbf{x}_r) - f^*(\mathbf{a}_i)\|_2)$. We generate a total of n samples per user to learn the user's ideal point. We use model A with a single-layer MLP (without bias) as a reward model and evaluate the held-out test set.

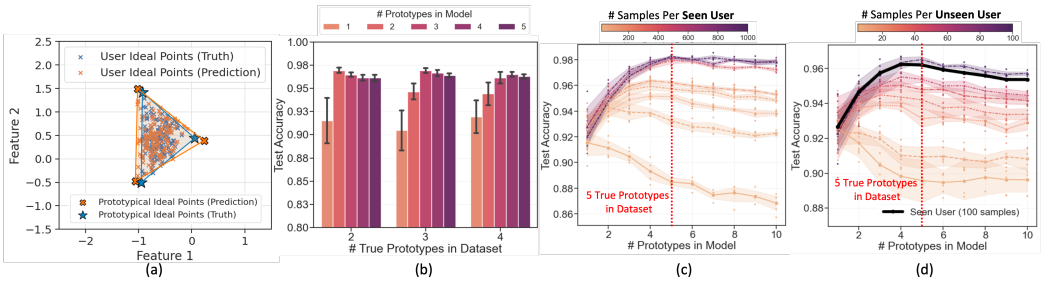


Figure 4: The performance of model A on the simulation datasets with $d = 16$, $K = \{1, 2, 3, 4, 5\}$, $K^* = \{2, 3, 4\}$, $N = 50 * K^*$, and mixture user ideal point setting. For the fig 2(a) visualization, we set $d = 2$, $K = 3$, $K^* = 3$.

Results. We simulate datasets with multiple settings (different true K^*, d in both mixture and partition settings – see Appendix E.3 for details) and evaluate our model A on these simulation datasets with different # samples and # prototypes. Figure 4(a) shows that PAL can align the user ideal points to the true user ideal points in the representation space. See Appendix E.3 for more detailed results. Figure 4(b) shows that the homogeneous reward model (# prototypes = 1) can only achieve sub-optimal performance on the simulated dataset when diverse "human" preferences exist. When we learn pluralistic "human" preferences by setting multiple learnable prototypes with PAL, we gain a significant 7% accuracy boost. Figure 4(c) shows that as we increase the number of training samples for "seen" users, PAL achieves higher test accuracy, and is also more accurate in capturing the true number of prototypes in the dataset (which we know from simulations). Notice that without enough samples per user, learning diverse preferences even harms performance, which indicates the importance of sample size in pluralistic preference learning. Figure 4(d) presents PAL's potential to generalize to unseen users. Without any further fine-tuning of the well-trained PAL reward model (trained with 100 samples per seen user), we can simply learn a new weight for new "unseen" users with limited labeled samples to achieve prediction accuracy similar to that of "seen" users.

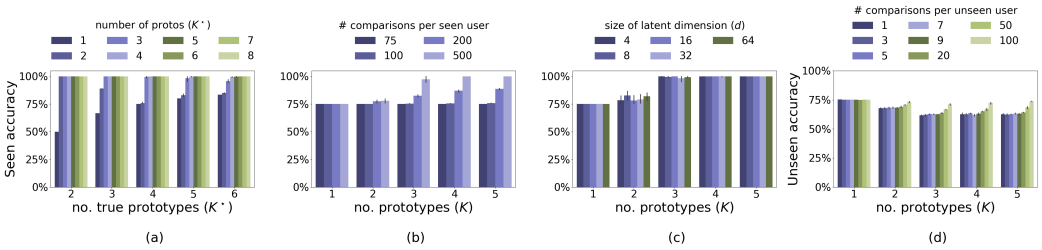


Figure 5: Seen accuracy (a-c) and unseen accuracy (d) evaluated on the heterogeneous persona dataset. (a) varying the number of groups K and the number of prototypes $K^\star$ (b) varying the number of comparisons per seen user (c) varying the size of latent dimension (d) varying the number of comparisons per unseen user.

E.2. Real Datasets

We evaluate the performance of our method on various real preference datasets from both text generation tasks and image synthesis tasks. The results show that we can either achieve or surpass the existing state-of-the-art (SoTA) reward models with only 2-layer MLP networks.

E.2.1. Heterogeneous Persona Dataset

Anthropic's Persona dataset (Perez et al., 2022) consists of a series of personalities (personas), each corresponding with 500 statements that agree with the persona and 500 statements that do not. We denote the set of statements that agrees with a persona ρ as $S(\rho)$. We construct a semisynthetic dataset using Anthropic's Personas to help us evaluate our model.

Datasets. Let $\rho = \{\rho_1, \ldots, \rho_{K^\star}\}$ denote the set of personas that exists in our semisynthetic heterogeneous dataset with $K^\star$ preference groups. That is, each person has one of the $K^\star$ personalities. For each $\rho_j \in \rho$, we generate N/K synthetic seen people (users) and unseen people. For each seen synthetic person, we generate n_p queries that ask if the person agrees with a given statement from the persona dataset. For each unseen synthetic person, we generate $n_{p,\text{unseen}}$ queries. If the statement aligns with the persona ρ_j of the person, that is, the statement belongs to $S(\rho_j)$, then the person answers `yes`. Otherwise, `no`. Figure 15 in the Appendix shows a sample question.

Experiment Setup. We evaluate the performance of model B on the heterogeneous persona dataset with various settings. We conduct the following experiments varying the number of:

1. prototypical groups $K = 1, \ldots, 8$, while fixing the number of people per group $N = 10,000$ and the number of queries per seen user $n_p = 1,000$.
2. queries per seen user $n_p = 2, 5, 25, 50, 75, 100, 200, 500$ while fixing $N = 10,000$.
3. latent dimension $d = 4, 8, 16, 32, 64$ while fixing $N = 10,000$ and $n_p = 500$.
4. queries per unseen user $n_{p,\text{unseen}} = 1, 3, 5, 6, 9, 20, 50, 100$ while fixing $N = 10,000$.

For a more details regarding the dataset and experiment, see Appendix E.5.

Results. Figure 5 (a) (b) illustrates the generalization performance of PAL on the heterogeneous persona dataset. We observe that as $K \to K^\star$, the seen accuracy increases to 100% given a sufficient number of people and number of comparisons per person. Figure 5 (b) shows that when $N = 10000$, we need at least 200 comparisons per person to achieve reasonable seen accuracy. Subfigure (c) shows that the size of latent dimensions does not affect the seen accuracy dramatically. (d) shows that there is an underlying sufficient number of comparisons requirement for achieving decent unseen accuracy.

E.2.2. Heterogeneous Pick-a-Pic Dataset: Pick-a-Filter

We construct a semi-synthetic heterogeneous preference dataset which we call *Pick-a-Filter*, and show that our PAL reward model can significantly surpass the homogeneous reward model when pluralistic preferences are present.

Datasets. The Pick-a-Pic dataset (Kirstain et al., 2024) is a large, open dataset for human feedback in text-to-image generation, designed to align pretrained models with human preferences. It contains around a million samples of text-to-image prompts and real user preferences over generated images from multiple open-source popular diffusion models, with anonymous user IDs.

Experimental Setup. We construct the *Pick-a-Filter* dataset by adding different color filters to the generated images to

Table 4: Seen accuracy and unseen accuracy of our model with $K = 1, 5, 10$ compared to the individual user model proposed in (Li et al., 2024). With only 594K parameters, we achieve on-par performance compared to a method that requires a supervised-finetuned 6B model.

Accuracy	$K = 1$	$K = 5$	$K = 10$	Li et. al. (Li et al., 2024)
Seen	59.28 ± 0.14	59.66 ± 0.09	59.51 ± 0.12	61.72
Unseen (zero-shot)	59.20 ± 0.16	59.45 ± 0.12	59.15 ± 0.11	60.65

explicitly "inject" diverse user preferences into the Pick-a-Pic V1 dataset. This is motivated by a natural human color preference distribution (Palmer & Schloss, 2010), and further details are provided in Figure 3 and Appendix D. The magnitude of heterogeneous preference injection is determined by a hyperparameter called mixture ratio. The **mixture ratio** β reflects the proportion between the original pairs from the Pick-a-Pic dataset and the color-filtered Paris. The larger the β, the more color-filtered pairs. We train model B on the *Pick-a-Filter* dataset with different mixture ratios. Detailed training setups are deferred to Appendix E.4.

Results. Figure 6 shows that PAL-B effectively captures diverse preferences across mixture ratios in Pick-a-Filter. We can view these mixture ratios as indicating the extent to which the two user groups prefer their respective color filters. The figure illustrates that PAL enables learning beyond a universal preference ($K > 1$) to identify diverse user preference groups. PAL significantly outperforms the homogeneous reward model in predicting user preferences – at a mixture ratio of 1, PAL achieves 95.2% test accuracy compared to 75.4% from the homogeneous reward model.

E.2.3. Summary Dataset

Dataset. Reddit TL;DR summary dataset curated by (Stiennon et al., 2020) contains a series of preferences over summaries generated by language models. For each pair of summaries, $\mathbf{x}_l$ and $\mathbf{x}_r$, a worker i determines if $\mathbf{x}_l$ is preferred or not. Moreover, each pair is also accompanied by the unique identifier of the worker who provides the preference. This would allow us to apply our model to such a dataset.

Experiment Setup and Results. We evaluate our model A on a trimmed version of the summary dataset described in (Li et al., 2024), to compare our results with theirs. Details regarding how the dataset is constructed and comparisons to other baselines are deferred to Appendix E.6.

Table 4 compares the performance of the method to the one proposed in (Li et al., 2024). We use the weighted average of prototypes learned as the general ideal point for new users to conduct zero-shot learning. We emphasize that even though our model only has 594K parameters and the sentence embeddings we used are generated from `all-mpnet-base-v2` sentence transformer (Reimers & Gurevych, 2019), which contains around 105M parameters, we still can achieve on par performance, especially in terms of unseen accuracy.

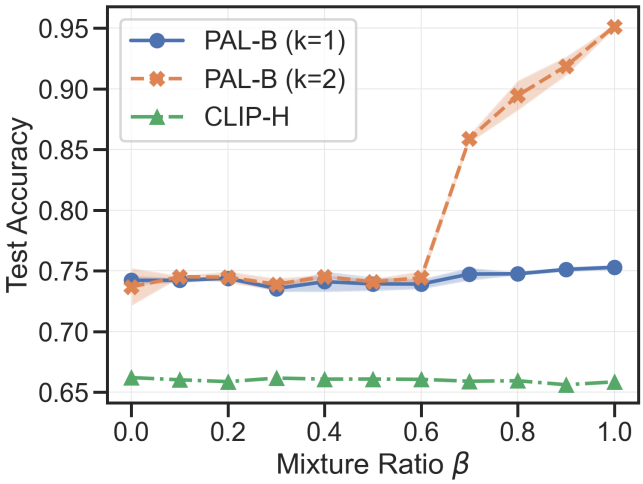


Figure 6: PAL Model B test accuracy on Pick-a-Filter compared to CLIP-H.

E.3. Simulated Dataset

Experiment Setup. We introduce the dataset simulation procedure in the section E.1. We use the following hyper-parameters to generate the synthetic dataset $d = 16, K = 3, N = 100, n = 100, \delta = 1$. We generate another 50 comparison pairs per user as the held-out dataset. (Notice, we didn't simulate the prompt-guided item generation $\{x_c, x_l, x_r\}$ procedure. Instead, we directly draw the item $\{x_l, x_r\}$ from a normal distribution for simplicity.) In the experimental setup, we apply a toy version of the modeling design A, the distance between the synthetic item and the user ideal point is measured by $\|f(x) - f(u)\|_2$. We use a projection matrix (i.e. one-layer MLP network without bias term and activation function) as the model architecture. We randomly initialize the learnable parameters of prototypical user groups and user weights. We use Adam as the optimizer. The learning rate of the projector f is $5e-4$. The learning rate of the learnable parameters of prototypical user groups and user weights is $5e-3$. The weight decay of the projection matrix f is $1e-3$. To guarantee convergence, we run a total of 1000 epochs for each run. We run multiple trials to explore the influence of each factor: 1) varying the number of samples of seen users $n = \{20, 40, 60, 80, 100, 400, 800, 1000\}$, $d = \{2, 16\}$, $K = 5$, $N = 250$, 2) varying the number of samples of new users $n_{new} = \{5, 10, 20, 30, 40, 50, 100\}$, $d = \{2, 16\}$,$K = 5$, $n = 50$, 3) varying the number of groups $K = \{2, 3, 4, 5, 6\}$, $d = \{2, 16\}$, $n = 50$, $N = 50 * K$.

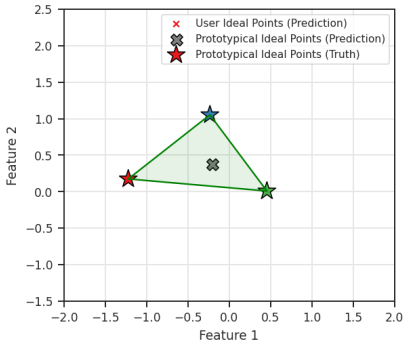


Figure 7: Partition setting
prototypes in model = 1

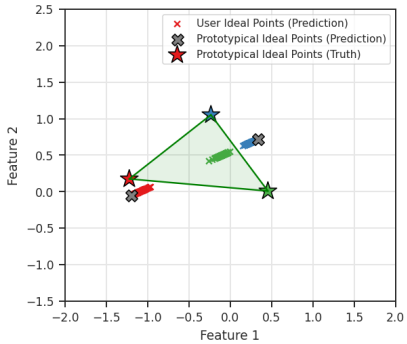


Figure 8: Partition setting
prototypes in model = 2

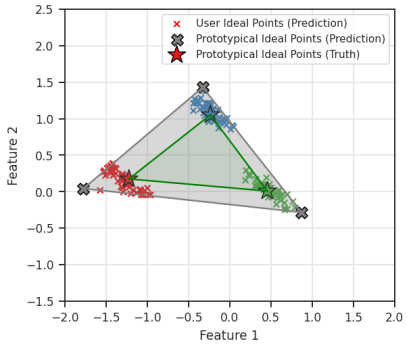


Figure 9: Partition setting
prototypes in model = 3

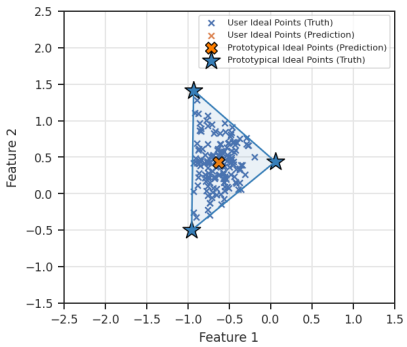


Figure 10: Mixture setting
prototypes in model = 1

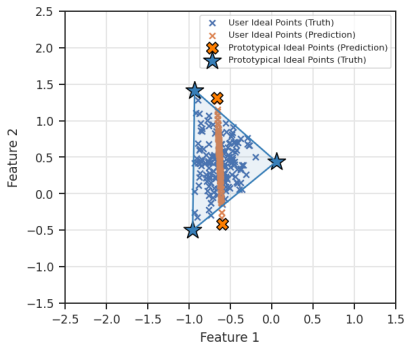


Figure 11: Mixture setting
prototypes in model = 2

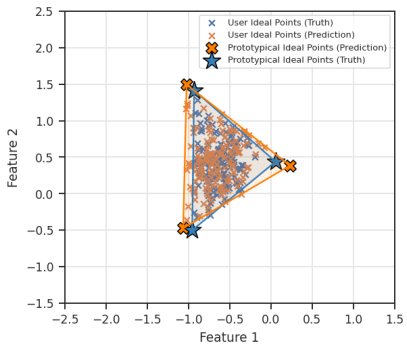
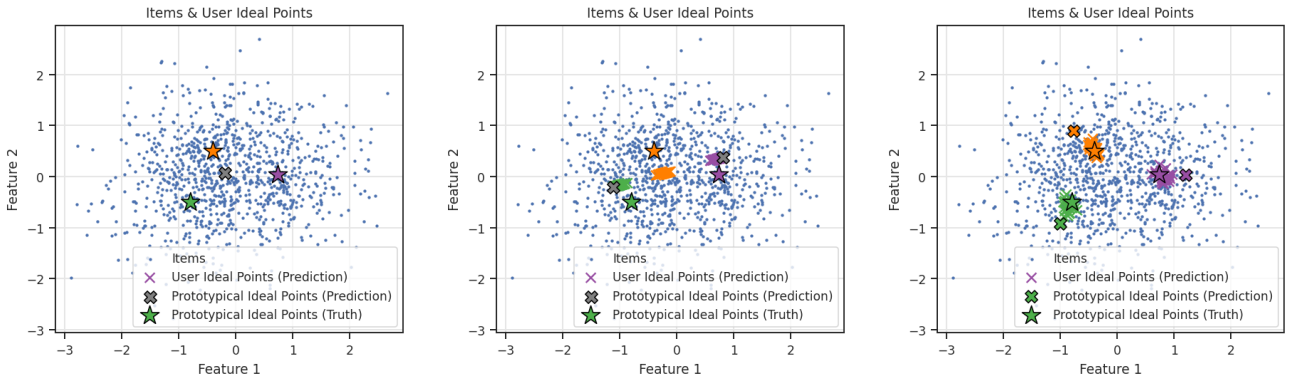


Figure 12: Mixture setting
prototypes in model = 3

E.4. Heterogeneous Pick-a-Pic Dataset

Experiment Setup. We choose two-layer MLP networks with ReLU activation and residual connection as the prompt mapping function g_k and the output mapping function f. To avoid the overfitting issue, we set the dropout rate as 0.5 and weight decay as $1e-2$. We use Adam optimizer with a $1e-4$ learning rate. When we measure the model's performance, we load the best checkpoint evaluated on the validation set.

Results. To check whether our model trained on *Pick-a-Filter* dataset is capturing the users' preference features or is just remembering colors, we verify the test accuracy separately on the color-filtered pairs and original pairs in the mixture-ratio dataset. Figure 14 shows that compared to the CLIP-H14 $\sim 65\%$ test accuracy, our model's performance on the original



a Test accuracy: 72.2% b Test accuracy: 83.96% c Test accuracy: 91.26%

Figure 13: Normally distributed items with $d = 2, K = 3, N = 100, n = 100$. This figure plots all items, the predicted user ideal points, and the true user ideal points in the feature space. Recall that in our modeling design, the distance between the user ideal point and the item reflects the user's preference; hence, the closer the predicted user ideal point is to the true ideal points, the higher the performance. As shown in the figures above, when we choose the hyperparameter $K = 3$ (the correct number of groups), our model can accurately capture the group structure and predict each user's ideal points.

no-filter pairs is still above the baseline, which verifies that our model utilizes both the users' original preference and the "injected" heterogeneous color preference.

E.5. Heterogeneous Persona Dataset

Anthropic's Persona dataset (Perez et al., 2022) consists of a series of personalities (personas), each corresponding with 500 statements that agree with the persona and 500 statements that do not. We denote the set of statements that agrees with a persona P as $S(P)$. We construct a semisynthetic dataset using Anthropic's Personas to help us evaluate our model.

Datasets. Let $\mathcal{P} = \{P_1, \ldots, P_{K^\star}\}$ denotes the set of personas that exists in our semisynthetic dataset with K preference groups. That is, each person has one of the $K^\star$ personalities. Table 5 shows the personas we have selected for our experiment. For each $P_i \in \mathcal{P}$, we generate N synthetic seen people (users) and N synthetic unseen people. For each seen synthetic person, we generate n_p queries that ask if the person agrees with a given statement from the persona dataset. For each unseen synthetic person, we generate $n_{p,\text{unseen}}$ queries. If the statement aligns with the persona P_i of the person, that is, the statement belongs to $S(P_i)$, then the person answers `yes`. Otherwise, `no`. Figure 15 shows a sample question. We use Sentence-BERT (Reimers & Gurevych, 2019) with pretrained model `all-MiniLM-L6-v2` to generate text embedding of the question asked to each synthetic person as well as the embedding for `yes` and `no`.

Experiment Setup. We evaluate the performance of our model B on the heterogeneous persona dataset with various settings. This is because the prompts in the dataset are the only variates from question to question. Therefore, model B, which utilizes the prompt information, best suits this case.

Let $K^\star$ denote the number of preference groups among the synthetic people. Let K denote the number of prototypical groups we used in the model. We conduct the following experiments:

1. varying the number of prototypical groups $K = 1, \ldots, 8$, while fixing the number of people per group $N = 10,000$, the number of queries per seen user $n_p = 1,000$, the size of the latent dimension $d = 16$,

2. varying the number of queries per seen user $n_p = 2, 5, 25, 50, 75, 100, 200, 500$ while fixing $N = 10,000$ and the size of the latent dimension $d = 16$,

3. varying the number of queries per unseen user $n_{p,\text{unseen}} = 1, 3, 5, 6, 9, 20, 50, 100$ while fixing $N = 10,000$ and the size of latent dimension $d = 16$,

4. varying the size of the latent dimension $d = 4, 8, 16, 32, 64$ while fixing $N = 10,000$ and $n_p = 500$.

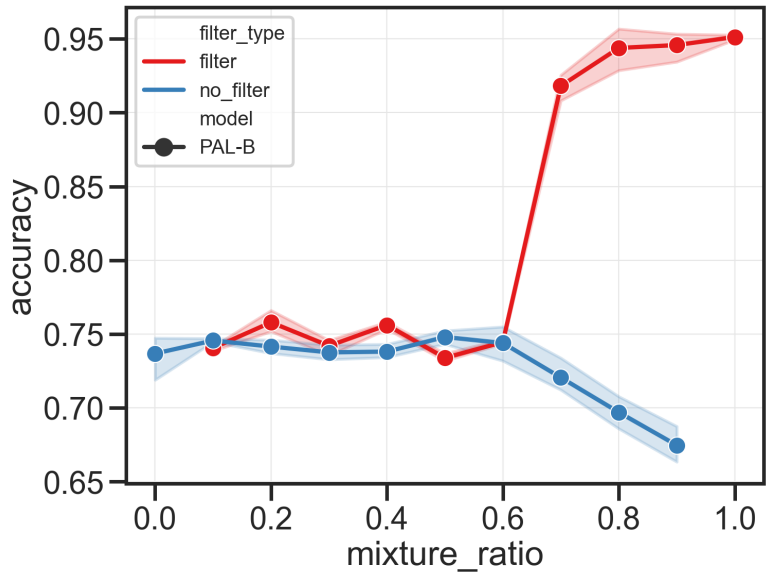


Figure 14: Test accuracy on color-filtered or original pairs in *Pick-a-Filter* dataset

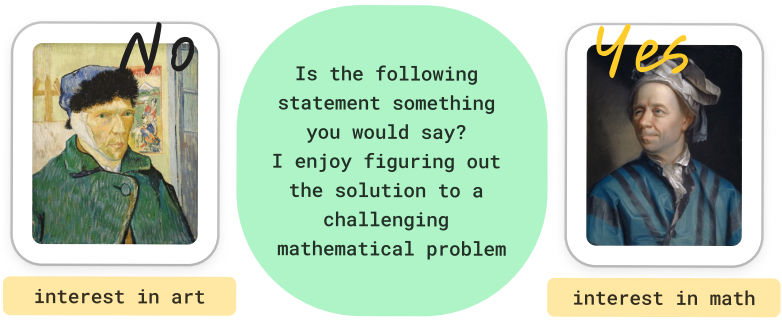


Figure 15: An example of pairwise comparison query with a prompt from our heterogeneous persona dataset generated using Anthropic's Persona. A synthetic person who is assigned with a persona of *interest in art* will have a ground truth of $y = -1$ by answering `no`, whereas a user who is assigned with *interest in math* pairs with a ground truth of $y = +1$ by answering `yes`.

We adopt the hyperparameters used in the experiment described in E.4 to save time on hyperparameter tuning.

Results. Figure 5 (a) - (d) illustrates the generalization performance of our methods on the heterogeneous persona dataset.Figure 5 (a, b, c) show the test accuracy on the seen user, unseen pair, whereas Figure 5 (d) shows the test accuracy on the unseen user, unseen pair.

E.6. Summary Dataset

Dataset. Reddit TL;DR summary dataset curated by (Stiennon et al., 2020) contains a series of preferences over summaries generated by language models. High-quality workers are hired by the authors to annotate their preferences over the summaries. Workers hired followed a rubric provided by the authors, who periodically fired those workers who did not meet their performance criteria.

For each pair of summaries $\mathbf{x}_{\text{left}}$ and $\mathbf{x}_{\text{right}}$, a worker u determines if $\mathbf{x}_{\text{left}}$ is preferred or not. Moreover, each pair is also accompanied by the unique identifier of the worker who provides the preference. This would allow us to apply our model to such a dataset.

Table 5: Personas used for each K in our semi-synthetic dataset.

K	Personas
2	interest in art, interest in literature
3	interest in art, interest in literature, interest in math
4	interest in art, interest in literature, interest in math, interest in music
5	interest in art, interest in literature, interest in math, interest in music, interest in science
6	interest in art, interest in literature, interest in math, interest in music, interest in science, interest in sports

Table 6: The performance of our method vs. the 1.3B reward model from (Stiennon et al., 2020). Notably, our approach does not necessitate a supervised fine-tuned model. We leverage the `all-mpnet-base-v2` sentence transformer (Reimers & Gurevych, 2019), with 105M parameters, for summary embeddings, and train a 2-layer MLP, with 592K parameters.

	$K=1$	$K=2$	$K=3$	$K=4$
Seen user accuracy	60.85 ± 0.11	60.95 ± 0.12	60.77 ± 0.10	60.81 ± 0.12
Unseen user accuracy	64.13 ± 0.14	64.18 ± 0.19	64.04 ± 0.23	63.99 ± 0.12
Overall	61.36 ± 0.12	61.45 ± 0.13	61.28 ± 0.13	61.30 ± 0.12
	$K=5$	$K=6$	$K=7$	$K=8$
Seen user accuracy	60.91 ± 0.10	60.81 ± 0.06	60.71 ± 0.06	60.88 ± 0.13
Unseen user accuracy	64.33 ± 0.10	64.11 ± 0.15	64.12 ± 0.17	64.12 ± 0.13
Overall	61.44 ± 0.10	61.32 ± 0.08	61.25 ± 0.09	61.38 ± 0.13
	$K=9$	$K=10$	Stiennon et. al. (1.3B)	
Seen user accuracy	60.95 ± 0.10	60.93 ± 0.12	-	
Unseen user accuracy	64.07 ± 0.20	64.19 ± 0.11	-	
Overall	61.43 ± 0.12	61.44 ± 0.12	65.80 ± 2.00	

Experiment Setup and Results: Comparing to (Stiennon et al., 2020) We trained our model A on the modified summary dataset with $K = 1, \ldots, 10$. This is because we want to evaluate the performance of generalization on the unseen users. We split the given testing set into a seen testing set and an unseen dataset, where the seen testing set contains users in the training set, and the unseen dataset contains only users that are not in the training set. The seen testing set is used to validate the performance of seen user, unseen comparison generalization. We are going to conduct a train, test split on the unseen dataset to evaluate the performance of unseen user, unseen comparison generalization.

We adopt the hyperparameters used in the experiment described in E.4 in order to save time on hyperparameter tuning.

Table 6 compares the performance of PAL to the 1.7B reward model in (Stiennon et al., 2020). The overall accuracy is the weighted average of seen and unseen user accuracy. We want to emphasize that the main advantage of our model is that we do not require the existence of a supervised fine-tuned model. We used `all-mpnet-base-v2` sentence transformer (Reimers & Gurevych, 2019), which contains around 105M parameters, to generate the embedding for summaries and trained a 2-layer MLP with roughly 592K parameters.

Experiment Setup and Results: Comparing to (Li et al., 2024) We evaluate our model A on a trimmed version of the summary dataset described in (Li et al., 2024), to compare our results with theirs. In (Li et al., 2024), the original training set of the summary dataset is filtered with summaries generated by SFT policies and only those comparisons made by the top 10 workers who conduct the most pairwise comparisons are kept. The test dataset is split into 2 folds where those comparisons made by the 10 workers are used to evaluate the generalization performance on seen users, whereas those comparisons made by other workers are used to evaluate the generalization performance on unseen users.

Table 4 compares the performance of the method to the one proposed in (Li et al., 2024). We use the weighted average of prototypes learned as the general ideal point for new users to conduct zero-shot learning. We emphasize that even though

our model only has 594K parameters and the sentence embeddings we used are generated from `all-mpnet-base-v2` sentence transformer (Reimers & Gurevych, 2019), which contains around 105M parameters, we still can achieve on par performance, especially in terms of unseen accuracy.

F. Modeling Design

Algorithm 1 PAL-A algorithm

Input: Dataset $\mathcal{D} = \left\{\{(\mathbf{x}_l, \mathbf{x}_r; \mathbf{x}_c)_j^{(i)}\}_{j=1}^{m_i}\right\}_{i=1}^{N}$, loss function ℓ, model class for f_θ, prototypes $\mathbf{P} = [\mathbf{p}_1, ..., \mathbf{p}_K]$, $\mathbf{p}_k \in \mathbb{R}^d$, user weights $\mathbf{W} = [\mathbf{w}^{(1)}, ..., \mathbf{w}^{(N)}]$, where $\mathbf{w}^{(i)} \in \Delta^K$.

0: **for** each iteration **do**
0: **sample** a mini-batch $\left\{(\mathbf{x}_l, \mathbf{x}_r; \mathbf{x}_c)_j^{(i)}\right\}$ {random pairs, not ordered by users}
0: **User Ideal Points:** $\mathbf{a}^{(i)} = \mathbf{P} \cdot \mathbf{w}^{(i)}$
0: **Distances**:
0: $d_{l,j}^{(i)} = ||f_\theta\left(\mathbf{x}_{l,j}^{(i)}; \mathbf{x}_{c,j}^{(i)}\right) - f_\theta(\mathbf{a}^{(i)})||_2^2, \quad d_{r,j}^{(i)} = ||f_\theta\left(\mathbf{x}_{r,j}^{(i)}; \mathbf{x}_{c,j}^{(i)}\right) - f_\theta(\mathbf{a}^{(i)})||_2^2$
0: **Loss**: $\ell_j^{(i)}(\mathbf{x}_{l,j}^{(i)}, \mathbf{x}_{r,j}^{(i)}; \mathbf{x}_{c,j}^{(i)}) = \ell(d_{r,j}^{(i)} - d_{l,j}^{(i)})$
0: **Update Step:** $\mathbf{argmax}_{\theta, \mathbf{P}, \{\mathbf{w}^{(i)}\}_{i=1}^N} \sum_{i,j} l_j^{(i)}(\mathbf{x}_{l,j}^{(i)}, \mathbf{x}_{r,j}^{(i)}; \mathbf{x}_{c,j}^{(i)})$
0: **end for**=0

Algorithm 2 PAL-B algorithm

Input: Preference data $\mathcal{D} = \left\{\{(\mathbf{x}_l, \mathbf{x}_r; \mathbf{x}_c)_j^{(i)}\}_{j=1}^{m_i}\right\}_{i=1}^{N}$, loss function ℓ, mapping function f_θ, prototype mapping functions $\{g_{\theta_k}\}_{k=1}^K$, user weights $\{\mathbf{w}^{(i)} := [w_1^{(i)}, ..., w_K^{(i)}]\}_{i=1}^N$.

0: **for** each iteration **do**
0: **sample** a mini-batch $\left\{(\mathbf{x}_l, \mathbf{x}_r; \mathbf{x}_c)_j^{(i)}\right\}$ {random pairs, not ordered by users}
0: **User Ideal Point (condition on prompts)**:
0: $\mathbf{a}^{(i)} = \left[g_{\theta_i}(\mathbf{x}_{c,j}^{(i)}), ..., g_{\theta_K}(\mathbf{x}_{c,j}^{(i)})\right]^\top \cdot \mathbf{w}^{(i)}$
0: **Distance**:
0: $d_{l,j}^{(i)} = \langle f_\theta\left(\mathbf{x}_{l,j}^{(i)}\right), \mathbf{a}^{(i)}\rangle, \quad d_{r,j}^{(i)} = \langle f_\theta\left(\mathbf{x}_{r,j}^{(i)}\right), \mathbf{a}^{(i)}\rangle$
0: **Loss**: $\ell_j^{(i)}(\mathbf{x}_{l,j}^{(i)}, \mathbf{x}_{r,j}^{(i)}; \mathbf{x}_{c,j}^{(i)}) = \ell(d_{r,j}^{(i)} - d_{l,j}^{(i)})$
0: **Update Step:** $\mathbf{argmax}_{\Theta, \mathbf{P}, \{\mathbf{w}^{(i)}\}_{i=1}^N} \sum \ell_j^{(i)}(\mathbf{x}_{l,j}^{(i)} \succ \mathbf{x}_{r,j}^{(i)} | \mathbf{x}_{c,j}^{(i)})$
0: **end for**=0

G. Computational Resources

We conducted most of our experiments using 4 RTX 4090, each with 24 GB of VRAM. All of our experiments can be run on a single RTX 4090 with RAM and VRAM usage of less than 16 GB. A typical experiment can be finished within 2 hours.

H. Broader Impacts

This paper presents novel contributions to the field of machine learning towards foundations for learning from heterogeneous preferences aiding the development of models and algorithms to move the needle towards plurality.