A Systematic Evaluation of Preference Aggregation in Federated RLHF for Pluralistic Alignment of LLMs

Mahmoud Srewa, Tianyu Zhao, & Salma Elmalaki
Department of Electrical Engineering and Computer Science
University of California, Irvine
Irvine, CA 92697, USA
{msrewa, tzhao15, salma.elmalaki}@uci.edu

Abstract

This paper addresses the challenge of aligning Large Language Models (LLMs) with diverse human preferences within Federated Learning (FL) environments where standard methods often fail to adequately represent diverse viewpoints. We introduce a comprehensive evaluation framework that systematically assesses the trade-off between alignment quality and fairness when using different aggregation strategies for human preferences. In our federated setting, each group locally evaluates rollouts and produces reward signals, and the server aggregates these group-level rewards without accessing any raw data. Specifically, we evaluate standard reward aggregation techniques (min, max, and average) and introduce a novel adaptive scheme that dynamically adjusts preference weights based on a group's historical alignment performance. Our experiments on Q/A tasks using a Proximal Policy Optimization (PPO)-based Reinforcement Learning from Human Feedback (RLHF) pipeline demonstrate that our adaptive approach consistently achieves superior fairness while maintaining competitive alignment scores. This work offers a robust methodology for evaluating Large Language Model (LLM) behavior across diverse populations and provides a practical solution for developing truly pluralistic and fairly aligned models.

1 Introduction

The remarkable capabilities of LLMs have positioned them as a central technology across various domains. However, their real-world utility and safety hinge on their ability to align with complex and diverse human values and social norms(1; 2). The prevailing methodology for this alignment is RLHF, which fine-tunes models based on collected human preference data (3). While effective, the standard RLHF paradigm often operates on a centralized dataset, which is not only a privacy concern but also risks embedding biases of a narrow demographic (4).

To address this, the integration of RLHF with FL has emerged as a promising avenue. FL allows for model training on decentralized data from numerous clients, thus preserving data privacy and capturing a wider range of human preferences (5; 6). However, this fusion presents a critical and underexplored challenge: **How to aggregate the diverse and potentially conflicting preference signals from different user groups?** In our setting, these preference signals appear as per-group reward scores generated locally by each client; the server aggregates these decentralized reward vectors without accessing any raw data to compute a global reward used for PPO updates. The choice of aggregation strategy is not merely a technical detail; it is an evaluation protocol that directly shapes the model's final behavior, determining whose preferences are prioritized and whose are marginalized.

This paper proposes a systematic evaluation framework to analyze the impact of different aggregation techniques on both alignment performance and fairness. By comparing standard reward aggregation methods with our proposed adaptive aggregation scheme, our goal is to define a more robust protocol

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Evaluating the Evolving LLM Lifecycle.

to assess LLMs in decentralized, pluralistic environments. We show that while simple aggregation methods can lead to unintended biases, our adaptive approach strikes a superior balance between achieving strong overall alignment and ensuring equitable representation across diverse groups, thus contributing to the development of more reliable and justly aligned LLMs. Our approach follows a zero-shot alignment paradigm, using only aggregated group reward signals without task demonstrations, ensuring generalizable alignment.

2 Background and Related Work

The alignment of LLMs with complex human preferences is a central goal in their development. Because explicitly encoding human values into a fixed loss function is challenging, *Reinforcement Learning from Human Feedback* (RLHF) has emerged as the dominant paradigm for steering models toward desired behavior. In RLHF, models learn from human preference data via a reward model trained on pairwise comparisons, and the policy is then optimized to maximize this learned reward (3). The most commonly used RL algorithm in RLHF is PPO, which fine-tunes the LLM using feedback from the reward model trained on human preference pairs (7). An alternative, Direct Preference Optimization (DPO), simplifies this pipeline by bypassing an explicit reward model and directly optimizing the policy to assign higher probability to preferred responses (8).

While traditional alignment methods typically aim for a single, global preference objective, real-world users exhibit diverse and sometimes conflicting preferences. Group Preference Optimization (GPO) (9) was introduced to address this challenge by enabling group-specific alignment of LLMs. GPO augments the base LLM with a lightweight transformer module trained via in-context supervised learning to predict and incorporate the distinct preferences of user groups using only a few examples. This auxiliary module serves as a preference predictor that captures alignment patterns across heterogeneous communities, allowing the model to dynamically adapt its responses according to different social or demographic contexts. In parallel, methods such as Group Robust Policy Optimization (GRPO) (10) and MaxMin-RLHF (11) have focused on ensuring robustness across diverse user populations within centralized RLHF settings. However, these methods still rely on collecting and processing user data on a central server, which raises privacy and data ownership concerns. To overcome these limitations,

To overcome these limitations, PluralLLM (6) extends GPO (9) into a federated learning architecture that enables groups to collaboratively learn lightweight transformer-based preference predictors without sharing raw data. Concretely, each group trains a lightweight transformer using few-shot in-context examples in a FL manner, producing a local preference module that can predict, for any Q/A question, a probability distribution over all answer options reflecting that group's latent preferences. This module serves as a fully local reward model; given RLHF rollouts from the server, the PluralLLM outputs group-specific preference probabilities that can be transformed into scalar rewards.

Our work builds directly on this foundation by using these PluralLLM predictors as decentralized reward generators for each group. At each PPO iteration, the server broadcasts rollouts to all groups, each group evaluates the responses using its PluralLLM module, and returns group-specific reward vectors. The central question we address is how to aggregate these heterogeneous and sometimes conflicting reward signals across groups. By systematically comparing standard aggregation schemes and introducing a dynamic alpha aggregation strategy, we analyze how different reward-aggregation protocols affect both fairness and alignment quality in multi-group federated RLHF.

3 Methodology

System Setup and Training Groups: In our setting, each group g_i corresponds to a distinct demographic or preference cluster (e.g., age, region, political leaning), and each group acts as a single federated client that locally represents its users' aggregated preferences. Our framework focuses on Q/A tasks with l training groups $G_{\text{train}} = \{g_1, g_2, \ldots, g_l\}$, where each group g_i maintains its private preference dataset $D_{g_i} = \{(x_{i,j}, y_{i,j})\}$ locally. Each preference sample consists of a query-response pair embedding $x_{i,j}$ and the corresponding group preference probability $y_{i,j}$. These datasets are distributed across groups and never shared with the central server, ensuring privacy preservation.

As illustrated in Figure 1, the aggregation server is initialized with a base LLM model π_θ^{base} and performs supervised fine-tuning (SFT) to adapt it for Q/A tasks, resulting in a policy model π_θ^{policy} suitable for PPO training. The server coordinates between policy optimization and distributed preference learning.

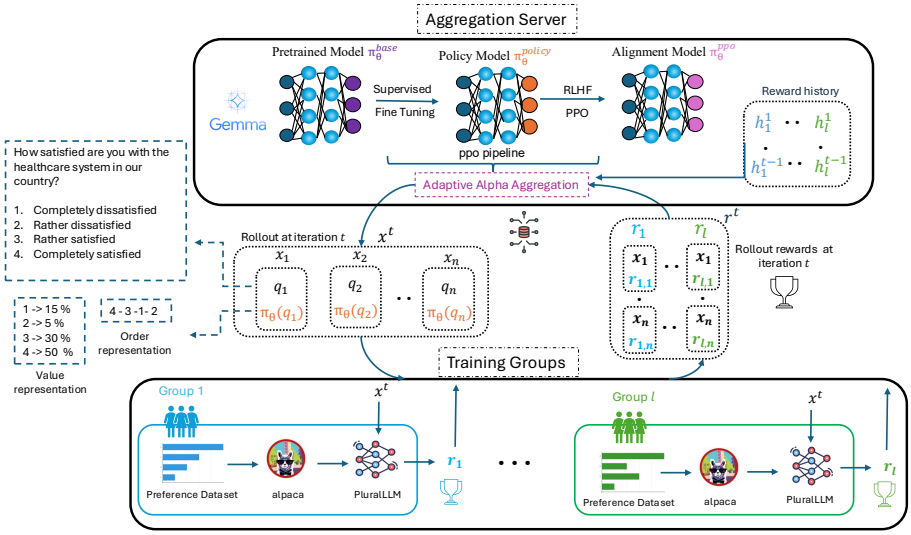

Figure 1: Federated RLHF for pluralistic alignment of group preferences in LLM.

At iteration t, the server generates rollouts X^t consisting of queries (questions with multiple choice's options) and LLM responses using the current policy π_θ^{policy}. These rollouts are distributed to all training groups for preference evaluation.

Distributed Reward Generation: Each group g_i first uses the PluraLLM (6) as a local lightweight reward model to generate preference probabilities for the received rollouts at iteration t. These probabilities are then converted to rewards $r_{g_i}^t$ according to the reward metric used. In our evaluation, we focus on two approaches: (1) preference probability prediction, where rewards are calculated directly from the predicted probabilities, and (2) preference ranking, where the probabilities are first converted to rankings before reward calculation. These task-specific rewards $r^t = \{r_{g_1}^t, r_{g_2}^t, \ldots, r_{g_l}^t\}$ reflecting how well the generated responses align with each group's preferences, are transmitted back to the aggregation server.

Adaptive Alpha Aggregation: The core of the federated RLHF framework is the aggregation of local rewards. At each training round, the central server receives per-group rewards r^t from different groups and aggregates them into a final global reward $Agg_\alpha(\mathbf{r}^t)$, ready to be used for updating the policy. We have $r^t = \{r_{g_1}^t, r_{g_2}^t, r_{g_3}^t, \ldots, r_{g_l}^t\}$ from l clients, where r_i^t represents how well the LLM outputs align with group g_i preference at FL iteration t. Recent work in the literature introduced an aggregation method, namely alpha aggregation(12) which achieves the consensus among heterogeneous feedback in RLHF. This is highlighted in Equation 1. This consensus reward aggregation is controlled by α, $Agg_\alpha(\mathbf{r}) = max(\mathbf{r})$, when $\alpha = \infty$ and $Agg_\alpha(\mathbf{r}) = min(\mathbf{r})$, when $\alpha = -\infty$.

$$Agg_\alpha(\mathbf{r}) = \begin{cases} \frac{1}{\alpha}\log(\frac{1}{N}\sum_{i\in N}\exp(\alpha r_i)) & \alpha \neq 0 \\ \frac{1}{N}\sum_{i\in N} r_i & \alpha = 0 \end{cases} \tag{1}$$

We propose an *adaptive* extension of this scheme that replaces the single global α with *group-specific*, dynamically updated weights α_g^t. Instead of using one fixed α for all clients, our method learns α_g^t per group based on its historical alignment performance, and plugs these into the alpha aggregation:

$$Agg_\alpha(\mathbf{r}^t) = \begin{cases} \frac{1}{|G_{\text{train}}|}\sum_{g\in G_{\text{train}}} r_g^t & \text{if } FI \geq 0.9 \\ \log\left(\frac{1}{|G_{\text{train}}|}\sum_{g\in G_{\text{train}}}\exp(\alpha_g^t \cdot r_g^t)\right) & \text{otherwise} \end{cases} \tag{2}$$

To achieve a balanced accumulated alignment reward history $\mathbf{h} = \{h_{g_1}, h_{g_2}, h_{g_3}, \ldots, h_{g_l}\}$, across clients, the aggregation weights α_i are dynamically adjusted in inverse proportion to each client's historical alignment performance($\alpha_i = \texttt{softmax}(1 - h_i)$). Specifically, a client i with a lower

accumulated alignment reward, h_i, is assigned a higher weight, α_i. As shown in Equation 2, a higher α_i value increases the dominance of the corresponding reward r_i. Hence, the α_i value for client i changes adaptively based on the alignment history for this client h_i.

Fairness Index (FI): To determine when uniform averaging is sufficient versus when adaptive weighting is needed, we use a FI that measures the similarity of per-group rewards $r^t_{g_i}$ across all groups. FI ranges from 0 to 1, where values near 1 indicate highly consistent (fair) rewards and lower values indicate increasing disparity. When $FI \approx 1$ (we use a practical threshold $FI \geq 0.9$), the rewards across groups are nearly uniform, and thus a simple average aggregation is appropriate. Otherwise, we employ the α-weighted log-sum-exp aggregation to amplify contributions from groups with historically lower alignment performance.

PPO Training and Iteration: With the aggregated rewards $Agg_\alpha(r^t)$, the server performs PPO optimization to update the policy model $\pi^{policy}_{\theta^t} \rightarrow \pi^{policy}_{\theta^{t+1}}$. The updated policy generates new rollouts for the next iteration, and the process continues until a predefined number of iterations or specific alignment score is reached. This iterative approach ensures continuous adaptation to diverse group preferences while maintaining fairness through our adaptive aggregation scheme.

4 Evaluation

We evaluate our approach using Gemma-2B-it, a fine-tuned version of the Gemma model, as our base LLM (13), More details on the experiment setup and configuration are summarized in Appendix A.

Our experiments utilize the Pew Research Center's Global Attitudes Surveys dataset (14), which captures diverse public opinions across social, political, and economic issues from various demographic groups. The dataset consists of multiple-choice survey questions answered by participants from a wide range of countries, with 2,554 questions spanning topics such as politics, media, technology, religion, race, and ethnicity. For each question and country, the dataset provides a probability vector over the answer choices, indicating the fraction of respondents in that country selecting each option. These probability vectors differ across countries, reflecting distinct group-level preferences. In our experiments, we treat each country as a separate user group (i.e., a federated client) and use all available groups in the survey. Our goal is to align the LLM with these diverse group preferences in a fair and robust manner, without overfitting to any single group or majority and without leaving any group behind.

We assess performance using fairness index FI and alignment scores across two primary tasks: the *preference probability prediction task* (see Figure 2) and the *preference ranking task* (see Figure 3). Our evaluation framework encompasses various reward functions and aggregation strategies. We compare our adaptive alpha aggregation against standard federated approaches (Min, Max, Average) and a supervised fine-tuning (SFT) baseline.

4.1 Evaluation Framework

The LLM rollout at FL iteration t, denoted X^t, consists of a set of questions $\{q_j\}$ together with corresponding responses generated by the policy model π_θ for each question q_j. We parse each response to extract the relevant output and denote the resulting LLM prediction as $y^{t,\text{llm}}_j$ for question j at iteration t. Let $o_{j,k}$ be the k-th answer option for question j, and let K denote the total number of answer options for that question. For each group $g \in G_{\text{train}}$, we compute a reward $r_{g,j}$ by comparing the LLM prediction $y^{t,\text{llm}}_j$ against the PluralLLM-derived target distribution for that group, $p^{\text{PluralLLM}}_{g,j}$ (6), i.e., the group-specific probability vector over options for question j predicted by the local PluralLLM reward model.

4.2 Reward Metrics

Our evaluation employs two categories of reward metrics to assess alignment quality across different aspects of preference modeling. These rewards are also used as alignment scores in our evaluation, i.e., they directly define the Avg AS and Min AS reported in Table 1.

4.2.1 Distance-Based Reward Metrics (Preference Prediction Task)

These rewards quantify the alignment between the LLM-predicted and PluralLLM-target probability distributions. They are computed locally on each client in G_{train}, where each group $g \in G_{\text{train}}$ independently evaluates the following reward functions:

Preference Probability Prediction Prompt

<bos><start_of_turn>user
You are an expert in modelling group preferences. You will receive **a question and exactly 4 options**.
Your task
- Assign a preference score to **each and every option**
- Produce **4** scores—no option may be skipped or combined
- Each score must be a decimal between 0 and 1, and **the rounded scores must sum to 1.00**
- Higher scores represent options a typical group is more likely to choose

Output format
- One line, comma-separated decimal numbers, **no spaces**
- Round each to **2** decimal places
- **No extra text, labels, or symbols**
- Example: 0.65,0.20,0.10,0.05

Return ONLY the 4 scores in the same order as options.
Question: Germany's influence in the EU Options: A: Has too much influence B: Has too little influence C: Has about the right amount of influence D: DK/Refused
<end_of_turn> <start_of_turn>model

Figure 2: Preference Probability Prediction Prompt

Preference Ranking Prompt

<bos><start_of_turn>user
You are an expert in ranking group preferences. You will receive **a question and exactly 4 options**.
Your task
- Rank all **4 provided options** from most to least preferred
- Process every option—no skipping or combining
- Order options based on what a typical group would most likely choose
- Higher preference options appear first

Output format
- One line, comma-separated option letters, **no spaces**
- Use the exact provided letters
- **No extra text, labels, or symbols**
- Example: B,C,A,D

Return ONLY the 4-letter ranking.
Question: Germany's influence in the EU Options: A: Has too much influence B: Has too little influence C: Has about the right amount of influence D: DK/Refused
<end_of_turn> <start_of_turn>model

Figure 3: Preference Ranking Prompt

Wasserstein Reward: Measures optimal transport cost between distributions.

$$r_{g,j}^{t,\text{Was}} = \frac{W_1(y_j^{t,\text{llm}}, p_{g,j}^{\text{PluralLLM}})}{K-1} \tag{3}$$

$r_{g,j}^{t,\text{Was}} \in [0,1]$, where 0 indicates a perfect distribution match. Lower values indicate better alignment between LLM and group preferences.

Cosine Similarity Reward: Captures directional similarity between preference vectors.

$$r_{g,j}^{t,\text{Cos}} = \frac{y_j^{t,\text{llm}} \cdot p_{g,j}^{\text{PluralLLM}}}{||y_j^{t,\text{llm}}|| \cdot ||p_{g,j}^{\text{PluralLLM}}||} \tag{4}$$

$r_{g,j}^{\text{Cos}} \in [-1,1]$, where 1 indicates identical direction, 0 indicates orthogonal, and -1 indicates opposite direction. Higher values indicate better preference alignment.

KL Divergence Reward: Measures information-theoretic alignment.

$$r_{g,j}^{t,\mathrm{KL}} = D_{\mathrm{KL}}\Big(p_{g,j}^{\text{PluralLLM}} \,\|\, y_j^{t,\text{llm}}\Big) = \sum_k p_{g,j,k}^{\text{PluralLLM}} \, \log \frac{p_{g,j,k}^{\text{PluralLLM}}}{y_{j,k}^{t,\text{llm}}}. \tag{5}$$

$r_{g,j}^{\text{KL}} \in [0, \infty)$, where 0 indicates identical distributions, and more positive = greater divergence. Smaller values indicate better alignment.

4.2.2 Ranking-Based Reward Metrics (Preference Ranking Task)

These rewards evaluate preference ordering consistency:

Kendall Tau Reward: Measures rank correlation between LLM and PluralLLM orderings.

$$r_{g,j}^{t,\text{Ken}} = \tau\Big(\text{rank}\Big(y_j^{t,\text{llm}}\Big), \text{rank}\big(p_{g,j}^{\text{PluralLLM}}\big)\Big). \tag{6}$$

$r_{g,j}^{t,\text{Ken}} \in [-1, 1]$, where 1 indicates perfect rank agreement, 0 indicates no correlation, and -1 indicates perfect disagreement. Higher values indicate better ranking alignment.

Borda Reward: Position-weighted scoring based on ranking accuracy.

$$r_{g,j}^{t,\text{Bor}} = \frac{\sum_{k=1}^{K}(K-k+1)\,\mathbb{I}\Big[\text{rank}\big(y_j^{t,\text{llm}}\big)_k = \text{rank}\big(p_{g,j}^{\text{PluralLLM}}\big)_k\Big]}{K(K+1)/2} \tag{7}$$

$r_{g,j}^{t,\text{Bor}} \in [0, 1]$, where 1 indicates perfect position-wise ranking match, and 0 indicates no correct positions. Higher values indicate better ranking quality.

Binary Reward: Simple correctness indicator.

$$r_{g,j}^{t,\text{Bin}} = \mathbb{I}[\text{rank}(y_j^{t,\text{llm}}) = \text{rank}(p_{g,j}^{\text{PluralLLM}})] \tag{8}$$

$r_{g,j}^{t,\text{Bin}} \in \{0, 1\}$, where 1 indicates exact ranking match, 0 indicates any disagreement. Binary indicator of perfect alignment.

4.3 Aggregation Schemes

For each question q_j at FL iteration t, the server aggregates the client-level rewards $\{r_{g_1,j}^t, r_{g_2,j}^t, \ldots, r_{g_l,j}^t\}$ across groups using different strategies:

Average Aggregation:

$$r_{\text{final},j}^t = \frac{1}{|G_{\text{train}}|} \sum_{g \in G_{\text{train}}} r_{g,j}^t \tag{9}$$

Provides balanced representation but may mask group-specific needs.

Min Aggregation:

$$r_{\text{final},j}^t = \min_{g \in G_{\text{train}}} r_{g,j}^t \tag{10}$$

Ensures no group is left behind but may be overly conservative, limiting overall performance.

Max Aggregation:

$$r_{\text{final},j}^t = \max_{g \in G_{\text{train}}} r_{g,j}^t \tag{11}$$

Optimizes for best-case performance but may neglect underrepresented groups.

Adaptive Alpha Aggregation:

$$r_{\text{final},j}^t = \begin{cases} \frac{1}{|G_{\text{train}}|} \sum_{g \in G_{\text{train}}} r_{g,j}^t & \text{if } FI \geq 0.9 \\ \log\left(\frac{1}{|G_{\text{train}}|} \sum_{g \in G_{\text{train}}} \exp(\alpha_g^t \cdot r_{g,j}^t)\right) & \text{otherwise} \end{cases} \tag{12}$$

Dynamically balances fairness and performance by favoring historically underperforming groups. The adaptive weights α_g^t are computed using reversed softmax on historical alignment scores:

$$\alpha_g^t = \frac{\exp((1 - h_g^{t-1})/T)}{\sum_{g' \in G_{\text{train}}} \exp((1 - h_{g'}^{t-1})/T)} \tag{13}$$

with temperature $T = 0.1$ and h_g^{t-1} being group g's historical alignment score.

Table 1: Fairness evaluation of pluralistic alignment across tasks, rewards, and aggregation strategies. FI denotes the Fairness Index. Alignment scores are reported under multiple metrics; higher is better for all metrics except KL and Was. Both average (Avg AS) and minimum (Min AS) alignment scores are shown. For each column, the best values are highlighted in **bold** (lowest for KL and Was, highest otherwise).

Task	Client Reward	Method	Server Agg.	Fairness Index (FI)						Avg Alignment Score (Avg AS)						Min Alignment Score (Min AS)					
				Was.	Cos.	KL	Ken.	Bor.	Bin.	Was.	Cos.	KL	Ken.	Bor.	Bin.	Was.	Cos.	KL	Ken.	Bor.	Bin.
Preference Prediction Task	—	SFT	—	0.98	0.97	0.88	0.85	0.83	0.97	0.10	0.82	0.4	0.28	0.38	0.23	0.08	0.77	0.55	0.13	0.34	0.23
	WassersteinReward	PPO	Alpha	**0.99**	**0.99**	**0.94**	**0.91**	**0.86**	**1.00**	0.05	0.90	0.26	0.30	0.42	0.22	**0.06**	**0.89**	**0.26**	0.21	0.37	0.22
			Min	0.98	**0.99**	0.93	0.87	0.79	0.90	0.05	**0.91**	**0.22**	0.42	0.44	0.27	**0.06**	**0.89**	0.27	0.34	0.41	**0.28**
			Avg	**0.99**	**0.99**	**0.94**	**0.91**	**0.86**	**1.00**	0.05	0.90	0.26	0.30	0.42	0.22	**0.06**	**0.89**	**0.26**	0.21	0.37	0.22
			Max	**0.99**	**0.99**	0.90	0.88	0.80	0.85	**0.03**	**0.91**	0.23	**0.45**	**0.51**	**0.31**	0.07	**0.89**	0.27	**0.43**	**0.47**	0.31
	CosineReward	PPO	Alpha	**0.99**	**0.99**	0.89	**0.88**	**0.89**	**0.91**	**0.05**	0.92	0.21	0.28	0.42	0.21	**0.06**	0.90	0.27	0.19	0.32	0.19
			Min	**0.99**	**0.99**	0.90	**0.88**	**0.89**	0.80	**0.05**	0.92	0.22	**0.34**	**0.45**	**0.28**	**0.06**	0.90	0.28	**0.21**	**0.34**	**0.22**
			Avg	**0.99**	**0.99**	0.89	**0.88**	**0.89**	**0.91**	**0.05**	0.92	0.21	0.28	0.42	0.21	**0.06**	0.90	0.27	0.19	0.32	0.19
			Max	**0.99**	**0.99**	**0.91**	0.87	0.88	0.88	**0.05**	**0.93**	**0.19**	0.31	0.42	0.22	**0.06**	**0.91**	**0.24**	0.20	**0.34**	0.19
	KLReward	PPO	Alpha	**0.99**	**0.99**	**0.92**	**0.92**	**0.90**	0.78	0.06	**0.92**	0.19	0.40	0.50	0.29	0.07	**0.90**	0.24	0.18	0.38	0.22
			Min	**0.99**	**0.99**	0.91	0.90	0.89	**0.84**	0.06	0.91	**0.17**	**0.43**	**0.51**	**0.33**	0.07	0.89	**0.22**	**0.33**	**0.43**	0.31
			Avg	**0.99**	**0.99**	0.91	0.89	**0.90**	0.76	0.05	0.91	0.19	0.40	0.48	0.27	0.06	0.89	0.26	0.29	0.40	**0.25**
			Max	**0.99**	**0.99**	0.91	0.90	0.86	0.75	**0.04**	0.91	0.19	0.33	0.40	0.20	**0.05**	0.88	0.24	0.22	0.33	0.19
	KendallTauReward	PPO	Alpha	**0.99**	**0.99**	**0.96**	0.90	0.71	**0.91**	0.07	0.75	0.48	0.43	0.38	**0.29**	0.08	0.72	0.55	0.34	0.36	**0.28**
			Min	**0.99**	**0.99**	0.95	0.90	0.75	**0.91**	0.07	0.73	0.49	**0.45**	**0.39**	**0.29**	0.08	0.71	0.54	0.37	0.36	**0.28**
			Avg	**0.99**	**0.99**	0.94	0.90	**0.76**	**0.91**	0.07	0.74	0.48	**0.45**	**0.39**	**0.29**	0.08	0.71	0.54	**0.38**	**0.37**	**0.28**
			Max	**0.99**	**0.99**	0.94	**0.92**	0.73	**0.91**	**0.06**	**0.76**	**0.44**	0.44	0.38	0.28	**0.07**	**0.73**	**0.50**	0.37	0.35	**0.28**
	BordaReward	PPO	Alpha	**0.99**	**0.99**	0.96	**0.89**	**0.71**	0.91	**0.08**	0.73	0.50	0.43	**0.39**	**0.29**	**0.09**	0.69	0.57	0.35	**0.36**	**0.28**
			Min	**0.99**	**0.99**	**0.98**	0.86	0.69	**0.92**	0.09	0.73	0.52	**0.44**	**0.39**	0.28	0.10	**0.70**	0.58	**0.37**	**0.36**	**0.28**
			Avg	**0.99**	**0.99**	0.97	**0.89**	**0.71**	0.91	0.09	0.73	0.51	0.42	0.38	**0.29**	0.10	**0.70**	0.58	0.35	**0.36**	**0.28**
			Max	**0.99**	**0.99**	0.97	**0.89**	**0.71**	0.90	**0.08**	**0.74**	**0.49**	0.44	**0.39**	**0.29**	**0.09**	**0.70**	**0.56**	0.36	**0.36**	**0.28**
	BinaryReward	PPO	Alpha	**0.99**	**0.99**	0.96	**0.89**	**0.68**	**1.00**	**0.08**	**0.74**	**0.52**	**0.42**	**0.38**	0.28	**0.10**	**0.70**	0.60	0.34	0.35	**0.28**
			Min	**0.99**	**0.99**	**0.98**	0.86	0.66	**1.00**	0.10	0.70	0.70	0.37	0.37	0.28	0.11	0.66	0.77	0.30	0.35	**0.28**
			Avg	**0.99**	**0.99**	0.97	**0.89**	0.67	**1.00**	0.09	0.73	0.54	0.42	**0.38**	**0.29**	**0.10**	**0.70**	**0.59**	**0.35**	**0.36**	**0.28**
			Max	**0.99**	**0.99**	0.97	**0.89**	**0.68**	**1.00**	**0.08**	0.73	0.56	0.41	**0.38**	0.28	**0.10**	**0.70**	0.64	0.33	0.34	**0.28**
Preference Ranking Task	—	SFT	—	—	—	—	0.89	0.87	0.83	—	—	—	0.38	0.50	0.31	—	—	—	0.25	0.41	0.27
	KendallTauReward	PPO	Alpha	—	—	—	**0.92**	0.81	0.97	—	—	—	**0.58**	0.47	**0.36**	—	—	—	**0.47**	0.42	**0.31**
			Min	—	—	—	**0.92**	0.81	**0.99**	—	—	—	0.52	0.46	0.30	—	—	—	0.43	0.41	0.28
			Avg	—	—	—	**0.92**	0.82	0.90	—	—	—	0.50	0.48	0.33	—	—	—	0.40	0.40	0.28
			Max	—	—	—	0.91	**0.88**	0.80	—	—	—	0.47	**0.53**	0.35	—	—	—	0.35	**0.44**	0.28
	BordaReward	PPO	Alpha	—	—	—	**0.94**	**0.95**	0.86	—	—	—	**0.53**	**0.61**	**0.39**	—	—	—	0.34	0.45	0.28
			Min	—	—	—	0.92	0.91	**0.89**	—	—	—	0.47	0.53	0.30	—	—	—	**0.36**	0.44	0.28
			Avg	—	—	—	0.93	0.92	0.86	—	—	—	0.49	0.58	**0.39**	—	—	—	0.35	**0.47**	**0.31**
			Max	—	—	—	0.91	0.92	0.78	—	—	—	0.45	0.54	0.32	—	—	—	0.34	0.45	0.28
	BinaryReward	PPO	Alpha	—	—	—	**0.91**	0.89	0.79	—	—	—	**0.49**	0.53	**0.35**	—	—	—	0.33	0.42	0.25
			Min	—	—	—	0.90	0.83	**0.90**	—	—	—	**0.49**	0.49	0.34	—	—	—	**0.39**	0.41	**0.28**
			Avg	—	—	—	**0.91**	0.90	0.79	—	—	—	**0.49**	0.53	**0.35**	—	—	—	0.37	**0.44**	**0.28**
			Max	—	—	—	**0.91**	**0.91**	0.79	—	—	—	0.47	**0.54**	**0.35**	—	—	—	0.35	**0.44**	**0.28**

4.4 Fairness Evaluation Metrics

The Fairness Index (FI) measures reward variation across groups for the same question-response pair:

$$FI = \frac{1}{|X|} \sum_{q_j \in X} \frac{1}{1 + \mathrm{CoV}^2(q_j)} \tag{14}$$

where the coefficient of variation for question q_j is:

$$\mathrm{CoV}(q_j) = \frac{\sigma\big(\{r^t_{g,j}\}_{g \in G_{\text{train}}}\big)}{\mu\big(\{r^t_{g,j}\}_{g \in G_{\text{train}}}\big)}. \tag{15}$$

$FI \in [0, 1]$, where 1 = perfect fairness (identical rewards across groups), and 0 = maximum unfairness. Higher FI values indicate more equitable treatment across demographic groups, while lower values suggest systematic bias favoring certain groups over others. We compute FI separately for each reward metric (Wasserstein, Cosine, KL, Kendall, Borda, Binary) to characterize fairness under different alignment criteria.

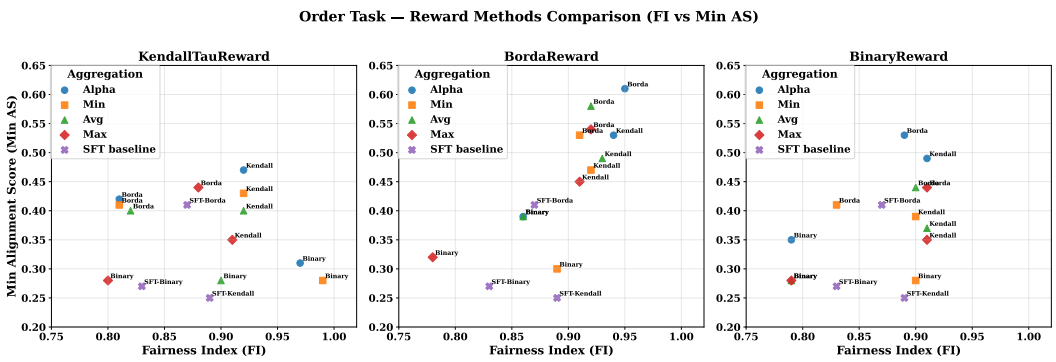

Figure 4: **Order task — FI vs. Min AS (worst–group performance) with reward-as-aggregation.** Each subplot fixes *one reward as the main aggregation metric* on the server— *KendallTauReward* (left), *BordaReward* (middle), and *BinaryReward* (right)— and then measures its effect on the three ranking metrics (Kendall, Borda, Binary). Points are server aggregation strategies (ALPHA, MIN, AVG, MAX); SFT baselines are purple crosses. We use the **minimum alignment score (Min AS)** on the y-axis because it reflects the performance of the *worst-served group*, while the x-axis shows the Fairness Index (FI).

4.5 Preference Probability Prediction Task Results and Analysis

The SFT baseline demonstrates suboptimal performance with fairness indices ranging from 0.83–0.98 and consistently lower alignment scores, highlighting the need for preference-based alignment. Detailed quantitative results are shown in Table 1 across multiple reward functions in the Value task.

Reward Function Analysis: Distance-based rewards (Wasserstein, Cosine, KL) substantially outperform ranking-based approaches (Kendall, Borda, Binary). Wasserstein and Cosine rewards with ALPHA aggregation achieve near-optimal fairness ($FI \approx 0.99$) while maintaining strong average alignment scores ($\text{Avg AS} \approx 0.90$–$0.95$). The minimum alignment scores, crucial for ensuring that no group is left behind, remain competitive ($\text{Min AS} \approx 0.89$–$0.94$), demonstrating an effective fairness–performance balance.

Dynamic Alpha Aggregation Strategy Impact: Across all distance-based rewards, our adaptive ALPHA aggregation consistently achieves superior fairness indices (often $FI = 0.99$) while preserving competitive average and minimum alignment scores. Compared to static strategies (MIN, AVG, MAX), ALPHA shows two key benefits: (i) it avoids the fairness degradation and worst-group collapse observed with MAX, which can push up average alignment at the cost of marginalized groups; and (ii) it improves worst-group performance relative to AVG, leading to higher Min AS at comparable or better FI. This behavior stems from dynamically reweighting groups based on historical alignment, allowing the server to upweight under-served groups without sacrificing overall utility.

Recommendations: For probability-based preference prediction tasks, we recommend using Wasserstein or Cosine rewards combined with ALPHA aggregation as the default configuration. This combination consistently achieves near-perfect fairness ($FI \approx 0.99$) while maintaining strong average and minimum alignment scores across groups. In settings where heightened sensitivity to distributional mismatch is desired, KL-based rewards with ALPHA aggregation remain competitive but may induce slightly larger fairness variance. Overall, Wasserstein/Cosine + ALPHA provides the most favorable fairness–performance trade-off for modeling calibrated group-level preference probabilities.

4.6 Preference Ranking Task Results and Analysis

The Order task evaluation focuses on ranking-based metrics (Kendall Tau, Borda, Binary). More details are shown in Table 1. The SFT baseline, trained on population-averaged preferences, shows particularly poor performance in ranking tasks with fairness indices of 0.83-0.89 and substantially lower alignment scores (0.31-0.50 average, 0.25-0.41 minimum). This performance degradation in ranking tasks underscores how averaged preference training fails to capture the nuanced ordering preferences that vary significantly across demographic groups.

Ranking Reward Analysis: Alpha aggregation with Kendall Tau rewards achieves the highest fairness index (0.92) and superior average alignment scores (0.58) compared to the SFT baseline (0.38). Borda rewards demonstrate the strongest overall performance, reaching fairness indices up to 0.95 with alpha aggregation and achieving the highest average alignment scores (0.61).

Figure 4 demonstrates that **ALPHA aggregation (blue circles) consistently provides the best fairness–performance trade-off,** occupying or approaching the upper-right quadrant ($\mathrm{FI} > 0.9$, $\mathrm{Min\,AS} > 0.3$) in all panels. For *KendallTauReward*, ALPHA attains high FI and strong worst-group performance across metrics (Kendall: $\mathrm{FI} \approx 0.92$, $\mathrm{Min\,AS} \approx 0.47$; Borda: $0.81, 0.42$; Binary: $0.97, 0.31$), whereas MIN pushes FI high on Binary (≈ 0.99) but *hurts* Min AS (≈ 0.28). For *BordaReward*, ALPHA achieves the *highest* Min AS across metrics (Kendall ≈ 0.53, Borda ≈ 0.61, Binary ≈ 0.39) with top/tied FI (Kendall ≈ 0.94, Borda ≈ 0.95, Binary ≈ 0.86). For *BinaryReward*, ALPHA again yields the largest Min AS (Kendall ≈ 0.49, Borda ≈ 0.53, Binary ≈ 0.35) with competitive FI, outperforming MIN/AVG/MAX in protecting the worst-served group. Across panels, the SFT baseline underperforms, reinforcing the need for federated preference alignment. Overall, the visualization shows that ALPHA most effectively resolves the fairness–performance tension by *maximizing worst-group (Min AS) performance at high FI.*

Dynamic Alpha Aggregation Strategy Impact: Across all ranking rewards, alpha aggregation maintains competitive performance while consistently achieving better fairness indices than alternative aggregation strategies. Importantly, our evaluation of minimum alignment scores reveals that alpha aggregation successfully prevents the marginalization of lowest-performing groups, maintaining minimum scores (0.31-0.47) that are competitive with or superior to other approaches, while simultaneously achieving higher average performance.

Recommendations: For order-based tasks, we recommend Borda rewards with alpha aggregation, which provides the optimal balance between fairness (0.94-0.95) and alignment performance (0.53-0.61 average). This combination effectively captures group ranking preferences while maintaining equitable treatment across demographic groups.

4.7 Overall Assessment

Our adaptive alpha aggregation demonstrates superior performance across both task types, consistently achieving the highest fairness indices while maintaining competitive alignment scores. The approach successfully addresses the critical challenge of preventing any group from being left behind, as evidenced by competitive minimum alignment scores across all evaluation scenarios. These results validate our hypothesis that adaptive weighting based on historical alignment performance provides an effective mechanism for achieving equitable federated learning in preference alignment tasks.

5 Limitations and Future Work

While our study demonstrates the effectiveness of adaptive alpha aggregation for pluralistic alignment, several areas present natural directions for further research.

Underlying RL Framework. Our current implementation relies on PPO. While effective, PPO can be computationally expensive. Future work should explore more resource-efficient alternatives such as GRPO or DPO, which would allow testing the aggregation strategy across different optimization paradigms and at larger scales.

Model and Dataset Scope. We evaluate on Gemma-2B-it using the Pew Research Global Attitudes dataset, which may be relatively conducive to cross-group alignment. Broader validation on base models and domains with more adversarial or conflicting preferences would provide a stronger stress test of the method's robustness.

Task Generalization. Our experiments focus on multiple-choice Q&A tasks, which offer a controlled setting for evaluation. Extending the framework to diverse tasks such as summarization, dialogue, or code generation would demonstrate its wider applicability and highlight how aggregation impacts more open-ended alignment scenarios.

These considerations do not detract from our main contribution—a systematic evaluation framework with a novel adaptive aggregation scheme—but rather open exciting avenues for expanding its applicability across models, datasets, and tasks.

6 Conclusion

This work tackles the challenge of aligning LLMs with diverse human preferences in decentralized, federated settings. We showed that *how* group feedback is aggregated is not a minor implementation detail, but a central part of the evaluation protocol that directly governs both fairness and overall alignment performance. Building on PluralLLM's group-specific preference modeling, we proposed an

evaluation framework that spans probability prediction and ranking tasks, multiple reward functions, and several aggregation baselines.

Within this framework, our Adaptive Alpha Aggregation dynamically reweights groups based on their historical alignment performance, consistently improving cross-group fairness while maintaining competitive alignment scores. In particular, it raises the performance of the worst-served groups without sacrificing overall utility, providing a practical path toward more pluralistic and equitably aligned LLMs in federated RLHF. This research contributes a valuable evaluation methodology to the field and opens up new avenues for future work, including applying this framework to a broader range of tasks and model architectures.

Acknowledgments

This work is supported by the U.S. National Science Foundation (NSF) under grant number 2339266.

References

[1] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, S. Zhong, B. Yin, and X. Hu, "Harnessing the power of llms in practice: A survey on chatgpt and beyond," *ACM Transactions on Knowledge Discovery from Data*, vol. 18, no. 6, pp. 1–32, 2024.

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

[3] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[4] S. Casper, X. Davies, C. Shi, T. K. Gilbert, J. Scheurer, J. Rando, R. Freedman, T. Korbak, D. Lindner, P. Freire *et al.*, "Open problems and fundamental limitations of reinforcement learning from human feedback," *arXiv preprint arXiv:2307.15217*, 2023.

[5] F. Wu, Z. Li, Y. Li, B. Ding, and J. Gao, "Fedbiot: Llm local fine-tuning in federated learning without full model," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 3345–3355.

[6] M. Srewa, T. Zhao, and S. Elmalaki, "Pluralllm: pluralistic alignment in llms via federated learning," in *Proceedings of the 3rd International Workshop on Human-Centered Sensing, Modeling, and Intelligent Systems*, 2025, pp. 64–69.

[7] J. Schulman *et al.*, "Proximal policy optimization algorithms," in *arXiv preprint arXiv:1707.06347*, 2017.

[8] R. Rafailov *et al.*, "Direct preference optimization: Your language model is secretly a reward model," *arXiv preprint arXiv:2305.18290*, 2023.

[9] S. Zhao, J. Dang, and A. Grover, "Group preference optimization: Few-shot alignment of large language models," *arXiv preprint arXiv:2310.11523*, 2023.

[10] S. S. Ramesh, Y. Hu, I. Chaimalas, V. Mehta, P. G. Sessa, H. Bou Ammar, and I. Bogunovic, "Group robust preference optimization in reward-free rlhf," *Advances in Neural Information Processing Systems*, vol. 37, pp. 37 100–37 137, 2024.

[11] S. Chakraborty, J. Qiu, H. Yuan, A. Koppel, F. Huang, D. Manocha, A. S. Bedi, and M. Wang, "Maxmin-rlhf: Alignment with diverse human preferences," *arXiv preprint arXiv:2402.08925*, 2024.

[12] C. Park, M. Liu, D. Kong, K. Zhang, and A. Ozdaglar, "Rlhf from heterogeneous feedback via personalization and preference aggregation," *arXiv preprint arXiv:2405.00254*, 2024.

[13] G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé *et al.*, "Gemma 2: Improving open language models at a practical size," *arXiv preprint arXiv:2408.00118*, 2024.

[14] E. Durmus, K. Nguyen, T. I. Liao, N. Schiefer, A. Askell, A. Bakhtin, C. Chen, Z. Hatfield-Dodds, D. Hernandez, N. Joseph *et al.*, "Towards measuring the representation of subjective global opinions in language models," *arXiv preprint arXiv:2306.16388*, 2023.

[15] L. von Werra, Y. Belkada, L. Tunstall, E. Beeching, O. Polleux, K. Rasul, L. Debut, and O. Sanseviero, "TRL: Transformer Reinforcement Learning," https://github.com/huggingface/trl, 2022.

A Experiment Configurations and Hyperparameters

Our experimental setup begins with supervised fine-tuning (SFT) as outlined in Table 2. We use the Gemma-2-2b-it model as our base, employing LoRA adaptation with rank 16 for efficient parameter updates. The SFT training utilizes a cosine learning rate scheduler with warmup and is conducted for a single epoch to establish our baseline model.

Table 2: SFT configuration and hyperparameters.

Hyperparameter	**Value**
Model	
Base model	`google/gemma-2-2b-it`
Precision	BF16
Data / Task	
Train/valid split	80/20%
Max sequence length	500 (include prompt and response)
LoRA Adapter	
Rank (r)	16
Alpha	32
Dropout	0.05
Optimization	
Batch size (per device)	16
Gradient accumulation steps	4
Learning rate	5×10^{-5}
Scheduler	cosine
Warmup steps	150
Weight decay	0.01
Training	
Epochs	1

As summarized in Table 3, both the policy and value models in PPO are initialized from the SFT model. During training, we employ two distinct prompt formats for evaluation: a preference probability prediction task requiring models to assign probability scores to all options, and a preference ranking task requiring complete ordinal ranking from most to least preferred (see Figures 2 and 3). Our implementation builds upon the Hugging Face TRL library (15). All experiments were conducted on 3 nodes, each equipped with A100 GPUs, Intel(R) Xeon(R) Gold 6326 CPUs @ 2.90GHz, and 256GB RAM.

[1]Rewards are whitened over each rollout before PPO updates.

Table 3: PPO configuration and hyperparameters (policy and value models initialized from the SFT model).

Hyperparameter	**Value**
General	
Policy model	Gemma 2 SFT model
Value model	Gemma 2 SFT model
Model / Quantization	
Quantization	4-bit (`nf4`, double-quant = True)
Compute dtype	BF16
Attention implementation	eager
LoRA (PEFT)	
Rank (r)	32
Alpha	32
Dropout	0.05
Optimization	
Per-device train batch size	4
Gradient accumulation steps	24
Learning rate	1×10^{-5}
Optimizer	AdamW
Weight decay	0.0
Scheduler	linear
PPO Trainer	
PPO epochs	2
Mini-batches	8
Per-device eval batch size	32
Response length	42
Temperature	0.6
KL coefficient	0.05
Clip range	0.2
Clip range (value)	0.2
Value loss coefficient (v_f)	0.2
Discount factor (γ)	1.0
GAE lambda (λ)	0.95
Reward whitening	Per rollout (before PPO update)[1]