FGD-Align: Pluralistic Alignment for Large Language Models via Fuzzy Group Decision-Making

Weihang Pan[1,2], Zhengxu Yu[2], Yong Wu[1], Xun Liang[3], Zhongming Jin[2], Qiang Fu[4], Penghui Shang[4], Binbin Lin[1*], Xiaofei He[3], Jieping Ye[2]

[1]School of Software Technology, Zhejiang University
[2]Alibaba Group
[3]State Key Lab of CAD&CG, Zhejiang University
[4]Hangzhou Zhiyuan Research Institute Co., Ltd
{panweihang, wu.yong, xunliang, binbinlin}@zju.edu.cn,
{yuzxfred, jinzhongming888, jieping}@gmail.com, xiaofeihe@cad.zju.edu.cn

Abstract

Ensuring alignment with human values is essential for modern large language models (LLMs), especially amid growing concerns around AI safety and social impact. Yet achieving such alignment remains challenging due to the limited, noisy, and often conflicting nature of human feedback from diverse annotators. Most existing approaches, such as Direct Preference Optimization (DPO), assume consistent and conflict-free supervision, overlooking the ambiguity, inconsistency, and value trade-offs inherent in real-world preferences—often leading to reduced robustness and exclusion of minority views. To address this, we propose **FGD-Align**, a novel pluralistic alignment framework grounded in Fuzzy Group Decision-Making theory. Our approach rigorously models and aggregates human preferences while retaining the complexity of real-world value trade-offs. Unlike traditional methods that rely on coarse-grained preference pairs, FGD-Align introduces fuzzy preference modeling via triangular fuzzy numbers to capture nuanced, multi-criteria human judgments. We further develop a new training objective, Probabilistic Fuzzy DPO, which incorporates fuzzy preference strength as adaptive loss weights and gradient filters, enhancing robustness to ambiguity and inconsistency in feedback. Comprehensive experiments demonstrate that FGD-Align consistently outperforms both DPO variants and advanced preference aggregation methods in terms of preference accuracy and robustness to ambiguity. It achieves superior alignment stability and better preserves minority preferences, all with minimal computational overhead. Our work bridges the gap between algorithmic tractability and the nuanced landscape of human values, enabling more scalable, inclusive, and socially-aware AI alignment.

Code — https://github.com/pwhjy/FGD-Align

Introduction

The rapid advancement of large language models (LLMs) and their widespread deployment across sensitive domains have amplified the urgency of ensuring alignment with broadly shared human values (Leike et al. 2018; Gabriel 2020). As LLMs increasingly influence decision-making, communication, and content generation, the challenge of aligning their behavior with the complex landscape of human ethics and social norms has emerged as a central concern in AI safety and governance (Dai et al. 2023).

Among existing alignment techniques, preference-based methods like Direct Preference Optimization (DPO) (Rafailov et al. 2023) and other Reinforcement Learning with Human Feedback (RLHF) (Zheng et al. 2023b) approaches have gained prominence due to their empirical effectiveness and conceptual simplicity. These methods leverage preference data (Song et al. 2024), typically in the form of pairwise comparisons, to steer model behavior toward preferred outputs. However, such approaches face a founda-

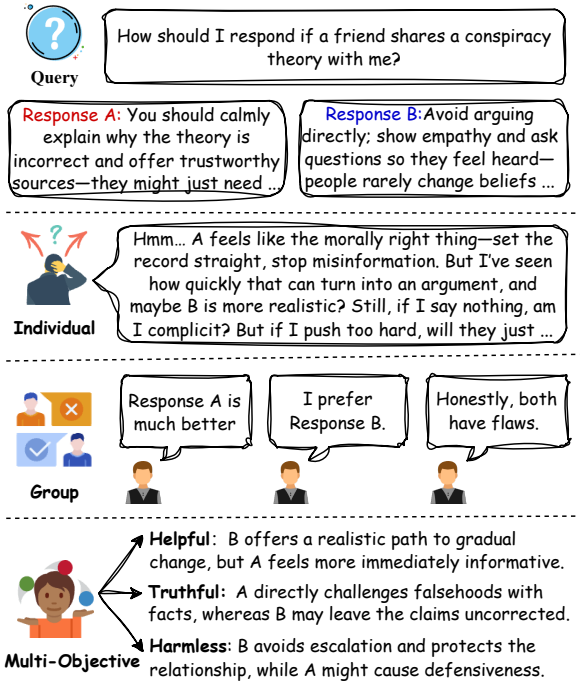

Figure 1: Illustration of three fundamental uncertainties in human preference annotation.

*Corresponding author

tional tension: they presume deterministic, conflict-free supervision signals, while real-world human preferences are inherently ambiguous, inconsistent, and multi-dimensional.

Most current alignment methods, including DPO, rely on the Bradley-Terry (BT) model (Hunter 2004), which compresses diverse and often conflicting human preferences into scalar scores to construct a "universal standard" (Sun, Shen, and Ton 2024). This simplification suffers from two key flaws: algorithmically, it biases the objective and undermines stability under conflicting feedback; societally, it risks marginalizing minority viewpoints, failing to capture the full diversity of human values. The BT model assumes clean, consistent annotations, a condition seldom met in practice due to the inherent ambiguity of human judgment. As shown in Figure 1, real-world preference data typically exhibits three forms of uncertainty: individual ambiguity, group-level inconsistency, and multi-objective conflict.

Recent works address these limitations through pluralistic alignment (Sorensen et al. 2024; Feng et al. 2024b), such as group-based reward modeling and personalized feedback learning, recognizing that human preferences reflect cultural, demographic, and ideological diversity and that a single perspective risks obscuring critical distinctions. Nonetheless, many such methods still lack principled mechanisms to account for individual ambiguity, intra-group inconsistency, and value conflicts across objectives.

In this paper, we propose **FGD-Align**, a modular framework for pluralistic alignment that addresses three core challenges in modeling human preferences: individual ambiguity, group inconsistency, and inter-criteria conflict. First, we introduce a fuzzy preference modeling module that maps linguistic and ordinal annotations into triangular fuzzy numbers (TFNs), allowing the model to represent uncertainty in subjective judgments. Second, we design a hierarchical aggregation scheme that combines robust intra-group consensus estimation with adaptive inter-criteria weighting based on annotator agreement, balancing majority signals with minority perspectives. Third, we extend the DPO objective into a probabilistic formulation that incorporates fuzzy consensus strength as soft supervision, enabling the model to modulate learning intensity based on preference certainty. Together, these components allow FGD-Align to capture the nuance, diversity, and uncertainty of real-world feedback, enhancing alignment accuracy, calibration, and value representation across benchmarks. Extensive experiments across multiple benchmark settings demonstrate that FGD-Align consistently outperforms strong baselines, including both DPO variants and preference aggregation methods, in terms of alignment accuracy, uncertainty calibration, and the preservation of diverse human perspectives.

Our contributions can be summarized as follows:

- We formulate the **Pluralistic Preference Alignment (PPA)** problem, which distinguishes between high-consensus and high-disagreement scenarios, and emphasizes the need for preserving viewpoint diversity under subjective or contested human feedback.
- We propose **FGD-Align**, a novel alignment framework based on fuzzy group decision-making, which integrates fine-grained fuzzy modeling of multi-criteria human preferences, a hierarchical aggregation scheme balancing consensus and minority opinions, and a probabilistic extension of DPO that dynamically adjusts learning to reflect confidence in human supervision.
- We introduce evaluation protocols tailored to the pluralistic setting, including **Accuracy** for consensus cases, **Probability Deviation Score** for uncertainty calibration in ambiguous samples, and **AUROC** for global separation quality. Extensive experiments show that FGD-Align outperforms existing DPO variants and preference aggregation methods across these axes.

Related Works

Early alignment methods (Shen et al. 2023; Wang et al. 2023b) train a reward model using human feedback, then apply reinforcement learning algorithms such as Proximal Policy Optimization (PPO) (Schulman et al. 2017) to adjust LLMs. Later, DPO removed the need for explicit reward modeling by aligning models directly on preference pairs. DPO variants (Saeidi et al. 2024) introduced finer control over alignment. However, these methods largely frame alignment as single-objective optimization, minimizing loss against a distilled representation of "human preference" from a small annotator pool (Feng et al. 2024a).

Recent literature increasingly critiques this "one-size-fits-all" paradigm (Xie et al. 2025), emphasizing that LLMs serve diverse populations with divergent values shaped by culture, demographics, and ideology (Artstein 2017; Plaza-del Arco et al. 2021). Jang et al. (Jang et al. 2023) argue that dominant objectives flatten variation, aligning toward an averaged preference. This "tyranny of the crowdworker" risks cultural homogenization and underrepresentation (Kirk et al. 2024a). In response, pluralistic alignment frameworks have emerged—approaches that recognize multiple valid preference sets rather than forcing consensus. These ideas draw from social choice theory (Sen 1986; Conitzer et al. 2024), employing aggregation methods like majority voting (Gasparin and Ramdas 2024), Borda counts, and Copeland scores to reconcile competing human values. Perspectivist approaches (Plank 2022; Akhtar, Basile, and Patti 2021; Muscato et al. 2025) further propose treating disagreement as a meaningful learning signal rather than collapsing labels, using hard and soft disaggregated annotations (van der Meer et al. 2024; Uma et al. 2022).

Evaluating pluralistic alignment remains a challenge. Traditional RLHF benchmarks (Zheng et al. 2023a; Bai et al. 2024) assume a single ground truth, whereas pluralistic alignment requires metrics that account for competing value systems. Recent work proposes preference-aware evaluation frameworks (Cheng et al. 2023; Zollo et al. 2024; Kumar et al. 2024; Salemi et al. 2024), including datasets like P-SOUP (Jang et al. 2023), which span multiple response styles, and PRISM (Kirk et al. 2024b), which embeds persona and demographic attributes into evaluations. Nonetheless, as Xie et al. (Xie et al. 2025) note, evaluation remains fragmented and heuristic-driven, lacking standardized protocols for measuring pluralism and fairness.

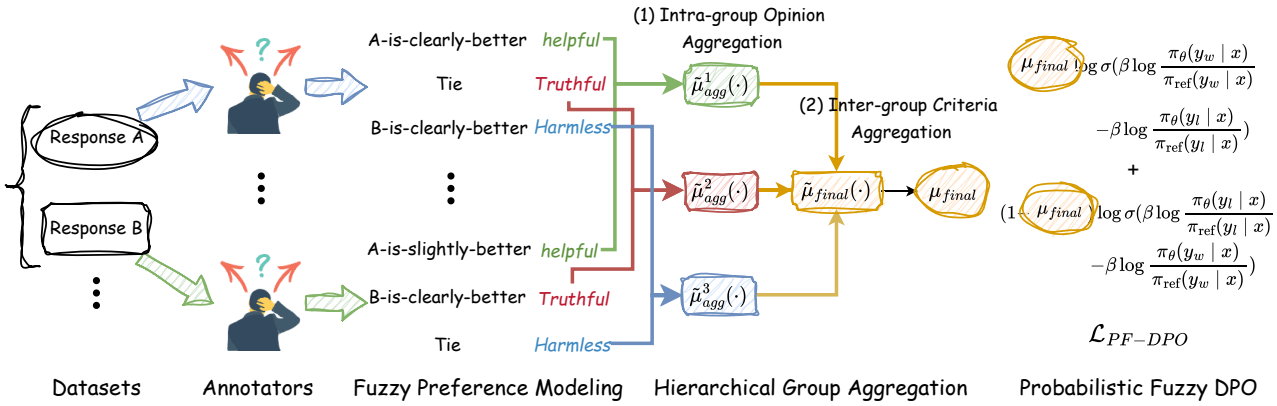

Figure 2: Overview of the FGD-Align framework. The framework comprises three coordinated components: (1) a fuzzy preference modeling module capturing judgment ambiguity, (2) a hierarchical group aggregation mechanism balancing conflict resolution with minority preservation, and (3) a probabilistic training objective adapting to preference certainty levels.

While prior work has modeled preference diversity via clustering, personalization, or voting-based aggregation, these methods often treat disagreement as noise or force it into discrete groups. In contrast, we propose a fuzzy decision-theoretic framework that embraces disagreement as a signal, enabling learning from both consensus and conflict.

Methodology

In this section, we first define the **Pluralistic Preference Alignment (PPA)** problem, then present our **FGD-Align** framework in detail.

Problem Definition

In the PPA setting, we consider a dataset containing N response pair instances associated with prompts $\{x_i\}_{i=1}^N$. Each instance $\{x_i, y_i^a, y_i^b, \{\{\mu_{i,k}^c\}_{c=1}^C\}_{k=1}^K\}_{i=1}^N$ consists of a prompt x_i, two responses y_i^a and y_i^b, and preference annotations $\mu_{i,k}^c$ from K users across C alignment criteria, where annotations may include relative preferences (e.g., "A-is-slightly-better") or Likert-scale ratings (e.g., [1-5]). The dataset intrinsically contains varying consensus levels, from high-agreement to high-disagreement cases.

The goal of PPA is dual-natured: For high-agreement data, the objective is to predict accurate binary preference labels $\hat{p}(y_i^a \succ y_i^b|x) \in \{0, 1\}$ identifying the objectively superior response. Conversely, for high-disagreement data, the objective shifts to avoiding strong preferences by assigning similar probabilities to both responses, thereby reflecting the ambiguity in human preference.

FGD-Align Framework

To effectively address the challenges posed by the PPA problem, as illustrated in Figure 2, we propose **FGD-Align**, a modular framework tailored to handle subjective, inconsistent, and multi-dimensional human preferences. In contrast to conventional approaches that reduce preferences to scalar labels or rely on simplistic majority voting, FGD-Align embraces the inherent fuzziness and diversity of real-world judgments by integrating three coordinated components.

First, we introduce a *fuzzy preference modeling* module to capture the linguistic uncertainty and subjective ambiguity in human annotations, enabling the system to represent nuanced opinions beyond scalar values. Second, we design a *hierarchical group aggregation* mechanism to robustly aggregate multiple viewpoints while preserving minority preferences, leveraging intra-group consistency and inter-criteria reliability to inform the preference aggregate. Finally, we develop a *probabilistic fuzzy DPO* objective that softens deterministic supervision by weighting learning according to consensus strength, thus adapting model behavior to both high-agreement and high-disagreement examples.

Together, these components form a unified solution that accommodates varying alignment, resolves conflicting preferences without discarding diversity, and guides model training with interpretable probabilistic signals. The following subsections detail the implementation of each module.

Fuzzy Preference Modeling We model subjective human preferences using a multi-criteria fuzzy decision system that transforms linguistic or numerical annotations into fuzzy representations. For each response pair (y_i^a, y_i^b) and alignment criterion c, individual annotations are converted into triangular fuzzy numbers (TFNs) $\tilde{\mu}_{i,k}^c = (l, m, u)$, where l, m, and u denote the pessimistic bound, core value, and optimistic bound, respectively. This representation captures the uncertainty and imprecision inherent in human judgment.

For annotations based on Likert-scale ratings (e.g., 1–5), we derive fuzzy preferences using the score difference $\Delta s = s(y_i^a) - s(y_i^b)$ between the two responses. The mapping from Δs to TFNs is defined as:

$$\tilde{\mu}_{i,k}^c = \begin{cases} \text{B-is-clearly-better} & (0.0, 0.1, 0.2) & |\Delta s| \leq -2 \\ \text{B-is-slightly-better} & (0.2, 0.3, 0.4) & |\Delta s| = -1 \\ \text{Tie} & (0.4, 0.5, 0.6) & \Delta s = 0 \\ \text{A-is-slightly-better} & (0.6, 0.7, 0.8) & |\Delta s| = 1 \\ \text{A-is-clearly-better} & (0.8, 0.9, 1.0) & |\Delta s| \geq 2 \end{cases}$$

This formulation enables the model to treat both ordinal annotations and soft linguistic labels in a unified fuzzy frame-

work. The resulting fuzzy preference vectors $\tilde{\mu}_{i,k}^{c}$ serve as the input for downstream aggregation and training modules.

Hierarchical Group Aggregation To integrate multiple annotators' preferences across diverse alignment criteria, we implement a hierarchical aggregation procedure consisting of intra-group and inter-group aggregation stages.

(1) Intra-group Opinion Aggregation. Within each criterion c, we aggregate annotators' fuzzy preferences $\{\tilde{\mu}_{i,k}^{c}\}_{k=1}^{K}$ into a robust group consensus using an iterative M-estimation process. Starting with the median TFN as the initial estimate $\tilde{\mu}_{\text{robust}}^{(0)}$, we iteratively reweight annotators based on their residual deviation from the current consensus:

$$d_k^{(t)} = \left|\tilde{\mu}_{i,k}^{c} - \tilde{\mu}_{\text{robust}}^{(t)}\right|^2,$$

$$\sigma^{(t)} = 1.4826 \times \text{MAD}(\{d_k^{(t)}\}),$$

$$w_k^{(t)} = \begin{cases} \left(1 - \left(\frac{d_k^{(t)}}{4.685\sigma^{(t)}}\right)^2\right)^2 & d_k^{(t)} \le 4.685\sigma^{(t)} \\ 0 & \text{otherwise} \end{cases}$$

Here, MAD stands for Median Absolute Deviation, which is less sensitive to outliers than standard deviation, making it suitable for robust estimation. The factor 1.4826 is used to make MAD a consistent estimator for the standard deviation under normality. The updated consensus is obtained by weighted averaging:

$$\tilde{\mu}_{\text{robust}}^{(t+1)} = \frac{\sum_{k=1}^{K} w_k^{(t)} \tilde{\mu}_{i,k}^{c}}{\sum_{k=1}^{K} w_k^{(t)}} \tag{1}$$

The process continues until convergence.

For each annotator, we compute a consistency score:

$$C_{i,k}^{c} = 1 - \left|\tilde{\mu}_{i,k}^{c}(y_i^a \succ y_i^b) - \tilde{\mu}_{\text{robust}}^{c}(y_i^a \succ y_i^b)\right| \tag{2}$$

Annotators are sorted by descending $C_{i,k}^{c}$, and their preferences are aggregated using Ordered Weighted Averaging (OWA) (Yager 1993). Given positional ratios $r_k = k/K$ and a fuzzy quantifier $Q(r) = \frac{1}{1+e^{-\beta(r-\alpha)}}$, where α and β control the shape of the sigmoid, we compute weights:

$$w_k = Q\left(\frac{r_k}{K}\right) - Q\left(\frac{r_k - 1}{K}\right) \tag{3}$$

The intra-criterion aggregated preference is then:

$$\tilde{\mu}_{\text{agg}}^{c}(y_i^a \succ y_i^b) = \sum_{k=1}^{K} w_k^{c} \cdot \tilde{\mu}_{k}^{c}(y_i^a \succ y_i^b) \tag{4}$$

(2) Inter-group Criteria Aggregation. To combine preferences across multiple criteria, we compute criterion-level reliability weights based on inter-annotator agreement. For each criterion c, we calculate the quadratic weighted Cohen's kappa (Warrens 2012; Wang et al. 2024):

$$\kappa_c = \frac{p_o - p_e}{1 - p_e}, \tag{5}$$

$$p_o = \frac{1}{K^2} \sum_{k=1}^{K} \sum_{k'=1}^{K} \left(1 - \frac{d_{kk'}^2}{d_{\max}^2}\right), \tag{6}$$

$$d_{kk'} = |\text{ord}(\mu_k^c) - \text{ord}(\mu_{k'}^c)| \tag{7}$$

Here, $\text{ord}(\mu_k^c)$ denotes the defuzzified scalar (e.g., centroid) of the fuzzy preference μ_k^c, used to compute the ordinal difference between annotators. d_{max} is the maximum possible ordinal difference for normalization. p_e is the expected agreement under random annotator behavior. The normalized reliability-based criterion weight is given by:

$$\omega_c = \frac{\kappa_c + \epsilon}{\sum_{c'=1}^{C}(\kappa_{c'} + \epsilon)}, \quad \epsilon = 10^{-8} \tag{8}$$

The final fused fuzzy preference is:

$$\tilde{\mu}_{\text{final}}(y_i^a \succ y_i^b) = \sum_{c=1}^{C} \omega_c \cdot \tilde{\mu}_{\text{agg}}^{c}(y_i^a \succ y_i^b) \tag{9}$$

Finally, we apply centroid defuzzification to obtain a scalar consensus strength:

$$\mu_{\text{final}} = \frac{l_{\text{final}} + m_{\text{final}} + u_{\text{final}}}{3} \tag{10}$$

where $l_{\text{final}}, m_{\text{final}}, u_{\text{final}}$ are the components of the final aggregated triangular fuzzy number $\tilde{\mu}_{\text{final}}$.

Probabilistic Fuzzy DPO We extend the Direct Preference Optimization (DPO) objective by incorporating fuzzy consensus strength into a probabilistic training formulation. Given a dataset of prompt-response triplets (x, y_w, y_l), where y_w is the preferred response and y_l the less preferred one, the standard DPO objective is defined as:

$$\mathcal{L}_{\text{DPO}} = -\,\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\sigma(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)})\Big] \tag{11}$$

This formulation implicitly assumes that all annotated preference pairs (y_w, y_l) exhibit deterministic preference relations, where y_w is strictly preferred over y_l (i.e., $p(y_w \succ y_l) = 1$). However, in real-world settings involving subjective human judgments, preferences are often uncertain or contested, and such hard supervision may lead to overconfident or biased learning.

To address this, we propose **Probabilistic Fuzzy DPO (PF-DPO)**, which generalizes the binary supervision by introducing a fuzzy preference strength $\mu_{\text{final}} \in [0, 1]$ derived from hierarchical aggregation. The modified objective is:

$$\mathcal{L}_{\text{PF-DPO}} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Bigg[\mu_{\text{final}} \log\sigma\left(\beta\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta\log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right) + (1-\mu_{\text{final}})\log\sigma\left(\beta\log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} - \beta\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)}\right)\Bigg] \tag{12}$$

This formulation enables the model to adapt its optimization strength to the degree of preference certainty. When μ_{final} approaches 1 or 0, it enforces confident directional learning; when $\mu_{\text{final}} \approx 0.5$, it softly balances between competing views, mitigating overfitting to noisy or ambiguous annotations.

Experiments

In this section, we comprehensively evaluate the proposed FGD-Align framework on three conflict-annotated datasets. Furthermore, we introduce novel metrics to quantify the preference divergence identified by the DPO models and perform an entropy-based analysis to gain deeper insights into the model's behavior.

Datasets. We utilize two human-annotated preference datasets and one LLM-annotated preference dataset to study diverging preferences: **Multipref** [1] (Zhang et al. 2024) contains 10K user-LLM preference pairs with fine-grained human annotations indicating the strength of preference. 39% of examples exhibit annotator disagreement ($\kappa = 0.268$), often caused by verbosity, task underspecification, and safety-related refusals. **HelpSteer2** [2] (Wang et al. 2024) includes 12K response pairs with 3–5 Likert-scale helpfulness ratings per sample. 24% show diverging preferences ($\kappa = 0.389$), with disagreements arising from verbosity, format, and subjective taste.

Multipref-LLMs is an LLM-annotated extension of Multipref, where four top-tier models (GPT-4-0409, Claude 3.5 Sonnet, Gemini 2.0 Flash, and Qwen-Max) act as annotators under multiple alignment criteria. More detailed dataset descriptions can be found in the Appendix.

For evaluation, we randomly sample 200 high-agreement and 200 high-disagreement instances from both the Multipref and HelpSteer2 datasets. However, the Multipref-LLMs dataset includes 200 high-agreement and 115 high-disagreement instances, as only 115 such cases were identified based on model annotation disagreement.

Metrics. We adopt distinct evaluation strategies for high-agreement and high-disagreement test sets: **High-Agreement Test Sets:** For samples where human preferences converge, we follow standard practices in reward model evaluation (Lambert et al. 2024) and use **Accuracy** for binary preference classification. This measures how well a model predicts the majority human preference when annotations are consistent. **High-Disagreement Test Sets:** For samples where human preferences diverge, we propose the **Probability Deviation Score (PDS)** to evaluate a model's ability to express uncertainty. Specifically, PDS measures how close the model's predicted preference probabilities are to 0.5, the ideal value for ambiguous cases:

$$\text{PDS} = 1 - \frac{1}{N_d} \sum_{i \in \mathcal{D}_d} |P(y_w \succ y_l \mid x_i) - 0.5| \qquad (13)$$

where $\mathcal{D}_d$ is the high-disagreement subset ($N_d = 200$). PDS ranges from 0 to 1, with higher values indicating better uncertainty calibration. The theoretical basis is that models should avoid overconfident predictions in the presence of subjective or conflicting human judgments. **AUROC (Across All Samples):** To further assess how well a model distinguishes between ambiguous and clear preference pairs, we compute the Area Under the Receiver Operating Characteristic Curve (AUROC). AUROC evaluates the model's ability to assign higher uncertainty (i.e., probabilities closer to 0.5) to disagreement-prone samples ($\mathcal{D}_d$), and more confident probabilities to agreement-prone samples ($\mathcal{D}_a$). Specifically, high-disagreement samples are treated as the positive class, and high-agreement samples as the negative class. AUROC is a threshold-independent metric ranging from 0.5 (random guessing) to 1.0 (perfect separation). It provides a holistic view of how well the model can identify and separate uncertain cases from reliable ones, regardless of whether predictions are ultimately correct.

Compared Methods. We systematically compare our approach with two primary categories of baselines: (1) **DPO Variants: Standard DPO** (Rafailov et al. 2023) is the fundamental preference optimization method, which directly optimizes language models using pairwise preference data; **KTO** (Kahneman-Tversky Optimization) (Ethayarajh et al. 2024) is a variant focusing on optimizing for human utility judgments using a Kahneman-Tversky theoretic loss; **SimPO** (Simple Preference Optimization) (Meng, Xia, and Chen 2024) is an efficient variant designed to maximize the likelihood of preferred responses with simplified objectives; **IPO** (Identity Preference Optimization) (Azar et al. 2024) introduces a KL regularization to DPO to mitigate overfitting and better preserve the model's original capabilities; and (2) **Preference Aggregation Methods: Majority Vote** (Wang et al. 2023a; Li et al. 2023) is the most commonly used strategy, selecting the option favored by the most annotators; **Social Choice Methods** (Dai and Fleisig 2024) adopt voting-theoretic approaches, such as Borda, which scores each option based on ranking positions. Condorcet, which seeks an option that wins all pairwise comparisons, and Copeland, which counts net wins across pairwise matchups, to more comprehensively aggregate individual preferences into a global ranking.

Implementation Details. For all methods, we train using 5 epochs with a warmup ratio of 0.1 and a cosine learning rate scheduler. Each model is trained using both learning rates: 2e-6 and 5e-6, and the preference loss weight is set to 0.1. Our proposed FGD-Align additionally introduces two hyperparameters, $\alpha = 0.5$ and $\beta = 10$. All experiments are conducted on 8 NVIDIA A100 GPUs.

Main Results

Performance Comparison Table 1 and Table 2 present a comprehensive evaluation of alignment methods on Qwen2.5-Instruct-7B across three preference datasets. All DPO Variants methods are trained using datasets constructed via Majority Vote aggregation, while our proposed method, FGD-Align, leverages fuzzy preference modeling through PF-DPO. From Table 1, we observe the following: 1) RewardBench leaderboard models achieve strong Preference Accuracy but suffer from lower Div.PDS scores, suggesting reduced robustness under annotator disagreement and potential overfitting to dominant signals. 2) FGD-Align consistently outperforms all baselines across all datasets and metrics, particularly excelling on disagreement-sensitive met-

[1] https://huggingface.co/datasets/allenai/multipref

[2] https://huggingface.co/datasets/nvidia/HelpSteer2/tree/main/disagreements

Method	Multipref			HelpSteer2			Multipref-LLMs		
	Pref.Acc	Div.PDS	AUROC	Pref.Acc	Div.PDS	AUROC	Pref.Acc	Div.PDS	AUROC
Representative Top-Ranked DPO Models from RewardBench									
stablelm-2-chat-12B	0.7950	0.8055	0.7449	0.9300	0.8091	0.7877	0.7800	0.7604	0.5512
Tulu-2-dpo-70B	0.7300	0.8954	0.6994	0.9150	0.8964	0.7623	0.7400	0.8218	0.5420
MMPO_Gemma-7B	0.8250	0.7210	0.7428	0.9500	0.7158	0.7944	0.8550	0.6795	0.5628
Baselines with DPO Variants									
Qwen2.5-Instruct-7B	0.7250	0.9199	0.7257	0.8900	0.9018	0.7918	0.7100	0.8759	0.5222
DPO	0.7950	0.8998	0.7506	0.9250	0.8827	0.8099	0.7850	0.8243	0.5410
KTO	0.8150	0.8824	0.7360	0.9450	0.8460	0.7911	0.7350	0.8548	0.5677
SimPO	0.8950	0.5974	0.7464	0.9550	0.5779	0.7517	0.8600	0.7447	0.5834
IPO	0.8400	0.7478	0.7742	0.9200	0.7330	0.7935	0.8200	0.8312	0.5319
FGD-Align w/ PF-DPO	**0.9000**	**0.9214**	**0.8119**	**0.9650**	**0.9216**	**0.8606**	**0.9000**	**0.8778**	**0.6491**

Table 1: Performance comparison of alignment methods on Qwen2.5-Instruct-7B across Multipref, HelpSteer2, and Multipref-LLMs. The table first lists representative top-ranked DPO models from the RewardBench leaderboard, followed by a comparison of various DPO variant methods (DPO, KTO, SimPO, IPO) trained using datasets constructed via majority voting. Best results are highlighted in bold, and second-best results are underlined.

Method	Multipref-LLMs		
	Pref.Acc	Div.PDS	AUROC
Majority Vote	0.7850	0.8243	0.5410
Borda	0.7850	0.8273	0.5474
Condorcet	0.7950	0.8253	0.5557
Copeland	0.8150	0.8258	0.5641
FGD-Align w/ DPO	**0.8550**	**0.8687**	**0.6004**

Table 2: Experiment results of different Preference Aggregation Methods. All datasets aggregated by these methods are trained using DPO. The Best results are highlighted in bold, while the second-best results are underlined.

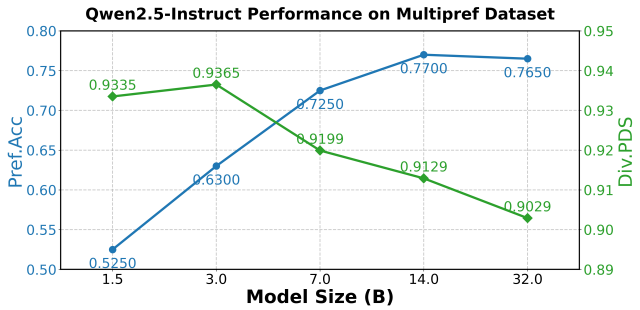

Figure 3: Performance comparison of Qwen2.5-Instruct models across different sizes.

rics (Div.PDS, AUROC), demonstrating better calibration and generalization under uncertainty. 3) SimPO achieves the second-best Preference Accuracy, which we attribute to its use of the sequence's mean log-probability as an implicit reward, mitigating length bias. However, due to the removal of the reference model and introduction of a target reward margin γ in the Bradley-Terry objective, SimPO tends to overoptimize, leading to substantially degraded Div.PDS scores.

Table 2 evaluates various preference aggregation strategies under a DPO framework, using the Multipref-LLMs dataset with complete pairwise annotations. Compared with classical aggregation methods, FGD-Align with fuzzy aggregation achieves the best overall results across all metrics. By modeling annotator consensus as a continuous, uncertainty-aware signal, it captures nuanced preferences more effectively and yields robust alignment performance.

Comparison of different model sizes Figure 3 illustrates the performance of Qwen2.5-Instruct models ranging from 1.5B to 32B parameters. As model size increases, Preference Accuracy improves steadily. However, Detection Accuracy declines, indicating a trade-off: larger models are more confident but less sensitive to ambiguous or uncertain inputs. This suggests potential overfitting to dominant preference patterns in larger models.

Additional experiments with Qwen3 and LLaMA 3.1 models are presented in the appendix.

Comparison of different epochs Figure 4 shows the training dynamics of DPO, SimPO, and FGD-Align from epochs 1 through 5. Preference Accuracy (left panel): SimPO increases rapidly and peaks at epoch 4. FGD-Align shows a steady upward trend, reaching the highest accuracy (0.9000) at epoch 5, highlighting its strong generalization. DPO improves until epoch 3 but then plateaus, suggesting limited optimization capacity. Probability Deviation Score (right panel): SimPO's PDS drops sharply (from 0.9037 to 0.5877), reflecting overconfidence and diminished sensitivity to ambiguous cases. DPO shows a gradual decline, stabilizing above 0.90. In contrast, FGD-Align maintains consistently high PDS (improving from 0.9185 to 0.9214), indicat-

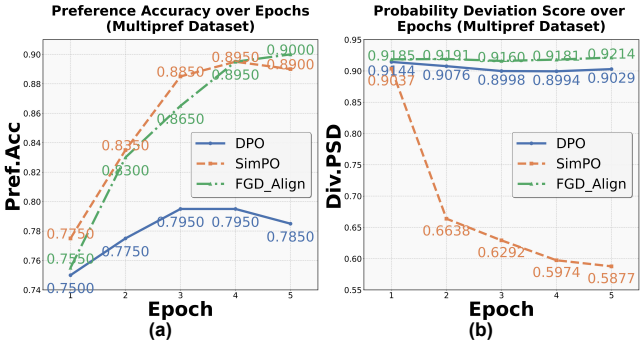

Figure 4: Performance of DPO variants on Multipref: Preference Accuracy (left) and Probability Deviation Score (right) over epochs.

Method	Multipref		
	Pref.Acc	Div.PDS	AUROC
FGD-Align w/ Single m	0.8550	0.9189	0.8040
FGD-Align w/ TFN	**0.9000**	**0.9214**	**0.8119**

Table 3: Ablation study comparing scalar-based and TFN-based annotator modeling for FGD-Align.

ing superior robustness and alignment with the underlying preference distributions.

Ablation Studies

Impact of Fuzzy Preference Modeling As shown in Table 3, modeling annotator preferences with a TFN yields consistently better results than using a single scalar value. The TFN-based variant improves accuracy, disagreement handling, and calibration, highlighting the advantage of capturing uncertainty in human annotations.

Effect of Consensus Strength Modeling As shown in Table 4, FGD-Align outperforms conservative DPO (cDPO) (Mitchell 2023), which applies uniform label smoothing with fixed μ_{final}. While cDPO treats all preferences as equally reliable, FGD-Align adaptively modulates learning strength based on fuzzy uncertainty, resulting in better preference accuracy, diversity preservation, and calibration.

Analysis From Entropy Perspective

To better understand the diversity of model outputs, we specifically select prompts that are labeled with annotator disagreement—a strong indicator that multiple valid answers may exist. Such prompts naturally require the model to preserve diversity in its responses to reflect the inherent ambiguity in the task. Table 5 reports the performance of three models—Qwen2.5-Instruct-7B, SimPO, and our proposed FGD-Align—across three core diversity metrics: 1) **Self-BLEU**: Measures the similarity between generated outputs. A higher Self-BLEU score indicates that the responses are more similar to each other, implying lower diversity. 2) **Distinct-2**: Quantifies the proportion of unique 2-grams in

Method	Multipref		
	Pref.Acc	Div.PDS	AUROC
cDPO($\mu_{\text{final}} = 0.1$)	0.7400	0.8739	0.5499
cDPO($\mu_{\text{final}} = 0.2$)	0.7100	0.8796	0.5543
FGD-Align	**0.9000**	**0.9214**	**0.8119**

Table 4: Comparison between fixed and fuzzy consensus strength modeling in PF-DPO. Conservative DPO uses fixed μ_{final} values, while FGD-Align adaptively adjusts learning strength via fuzzy preference uncertainty.

Model	Self-BLEU ↓	Distinct-2 ↑	Entropy ↑
Base Model	0.6963	0.2923	6.6575
DPO	0.7356	0.2263	5.7523
SimPO	0.9875	0.0066	1.3275
FGD-Align	**0.5719**	**0.4376**	**6.8743**

Table 5: Diversity evaluation on prompts with annotator disagreement. The Base Model is Qwen2.5-7B-Instruct, while all other methods are fine-tuned on the Multipref dataset.

the generated texts. Higher values reflect richer lexical variation. 3) **Entropy**: Computes the Shannon entropy of the sampled response distribution, representing the overall uncertainty and spread of model outputs. A lower entropy suggests the model collapses to a narrow set of responses.

As shown in the table, FGD-Align consistently demonstrates stronger diversity across all metrics. Specifically, it achieves lower Self-BLEU scores, higher Distinct-2 values, and higher output entropy compared to SimPO. These results indicate that FGD-Align is better able to preserve output diversity on prompts where multiple valid answers exist, mitigating the risk of mode collapse. In contrast, SimPO shows signs of over-optimization, generating overly similar responses with reduced entropy.

Conclusion

In this paper, we propose FGD-Align, a novel pluralistic alignment framework for large language models grounded in Fuzzy Group Decision-Making theory. Our method models nuanced, multi-criteria human preferences using triangular fuzzy numbers and introduces a hierarchical aggregation mechanism to balance consensus with diversity. We further develop Probabilistic Fuzzy DPO, a new training objective that adaptively incorporates fuzzy preference strength, improving robustness to ambiguity and inconsistency. Experiments across three datasets demonstrate that FGD-Align outperforms strong baselines in preference accuracy, stability, and minority preservation. Despite its effectiveness, our current work is limited by the scale and diversity of human annotations. In future work, we aim to incorporate broader annotator populations from online communities and explore more expressive fuzzy representations to better capture complex human value judgments.

Acknowledgments

This work was supported in part by the National Nature Science Foundation of China (Grant No: 62273303), in part by Yongjiang Talent Introduction Programme (2022A-240-G), in part by Ningbo Key R&D Program (2023Z229).

References

Akhtar, S.; Basile, V.; and Patti, V. 2021. Whose Opinions Matter? Perspective-aware Models to Identify Opinions of Hate Speech Victims in Abusive Language Detection. arXiv:2106.15896.

Artstein, R. 2017. *Inter-annotator Agreement*, 297–313. Dordrecht: Springer Netherlands. ISBN 978-94-024-0881-2.

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

Bai, G.; Liu, J.; Bu, X.; He, Y.; Liu, J.; Zhou, Z.; Lin, Z.; Su, W.; Ge, T.; Zheng, B.; et al. 2024. MT-Bench-101: A Fine-Grained Benchmark for Evaluating Large Language Models in Multi-Turn Dialogues. *arXiv preprint arXiv:2402.14762*.

Cheng, P.; Xie, J.; Bai, K.; Dai, Y.; and Du, N. 2023. Everyone Deserves A Reward: Learning Customized Human Preferences. arXiv:2309.03126.

Conitzer, V.; Freedman, R.; Heitzig, J.; Holliday, W. H.; Jacobs, B. M.; Lambert, N.; Mossé, M.; Pacuit, E.; Russell, S.; Schoelkopf, H.; et al. 2024. Social choice should guide ai alignment in dealing with diverse human feedback. *arXiv preprint arXiv:2404.10271*.

Dai, J.; and Fleisig, E. 2024. Mapping Social Choice Theory to RLHF. arXiv:2404.13038.

Dai, J.; Pan, X.; Sun, R.; Ji, J.; Xu, X.; Liu, M.; Wang, Y.; and Yang, Y. 2023. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*.

Ethayarajh, K.; Xu, W.; Muennighoff, N.; Jurafsky, D.; and Kiela, D. 2024. KTO: Model Alignment as Prospect Theoretic Optimization. arXiv:2402.01306.

Feng, D.; Qin, B.; Huang, C.; Zhang, Z.; and Lei, W. 2024a. Towards analyzing and understanding the limitations of dpo: A theoretical perspective. *arXiv preprint arXiv:2404.04626*.

Feng, S.; Sorensen, T.; Liu, Y.; Fisher, J.; Park, C. Y.; Choi, Y.; and Tsvetkov, Y. 2024b. Modular pluralism: Pluralistic alignment via multi-llm collaboration. *arXiv preprint arXiv:2406.15951*.

Gabriel, I. 2020. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3): 411–437.

Gasparin, M.; and Ramdas, A. 2024. Merging uncertainty sets via majority vote. *arXiv preprint arXiv:2401.09379*.

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

Jang, J.; Kim, S.; Lin, B. Y.; Wang, Y.; Hessel, J.; Zettlemoyer, L.; Hajishirzi, H.; Choi, Y.; and Ammanabrolu, P. 2023. Personalized Soups: Personalized Large Language Model Alignment via Post-hoc Parameter Merging. arXiv:2310.11564.

Kirk, H. R.; Vidgen, B.; Röttger, P.; and Hale, S. A. 2024a. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 6(4): 383–392.

Kirk, H. R.; Whitefield, A.; Röttger, P.; Bean, A.; Margatina, K.; Ciro, J.; Mosquera, R.; Bartolo, M.; Williams, A.; He, H.; Vidgen, B.; and Hale, S. A. 2024b. The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. arXiv:2404.16019.

Kumar, I.; Viswanathan, S.; Yerra, S.; Salemi, A.; Rossi, R. A.; Dernoncourt, F.; Deilamsalehy, H.; Chen, X.; Zhang, R.; Agarwal, S.; et al. 2024. Longlamp: A benchmark for personalized long-form text generation. *arXiv preprint arXiv:2407.11016*.

Lambert, N.; Pyatkin, V.; Morrison, J.; Miranda, L.; Lin, B. Y.; Chandu, K.; Dziri, N.; Kumar, S.; Zick, T.; Choi, Y.; et al. 2024. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*.

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

Li, Y.; Lin, Z.; Zhang, S.; Fu, Q.; Chen, B.; Lou, J.-G.; and Chen, W. 2023. Making Large Language Models Better Reasoners with Step-Aware Verifier. arXiv:2206.02336.

Meng, Y.; Xia, M.; and Chen, D. 2024. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37: 124198–124235.

Mitchell, E. 2023. A Note on DPO with Noisy Preferences & Relationship to IPO. https://ericmitchell.ai/cdpo.pdf. Accessed: 2024-07-13.

Muscato, B.; Li, Y.; Gezici, G.; Zhao, Z.; and Giannotti, F. 2025. Bridging the Gap: In-Context Learning for Modeling Human Disagreement. *arXiv preprint arXiv:2506.06113*.

Plank, B. 2022. The 'Problem' of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. arXiv:2211.02570.

Plaza-del Arco, F. M.; Montejo-Ráez, A.; Ureña-López, L. A.; and Martín-Valdivia, M.-T. 2021. OffendES: A New Corpus in Spanish for Offensive Language Research. In Mitkov, R.; and Angelova, G., eds., *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, 1096–1108. Held Online: INCOMA Ltd.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36: 53728–53741.

Saeidi, A.; Verma, S.; Uddin, M. N.; and Baral, C. 2024. Insights into alignment: Evaluating dpo and its variants across multiple tasks. *arXiv preprint arXiv:2404.14723*.

Salemi, A.; Mysore, S.; Bendersky, M.; and Zamani, H. 2024. LaMP: When Large Language Models Meet Personalization. arXiv:2304.11406.

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

Sen, A. 1986. Social choice theory. *Handbook of mathematical economics*, 3: 1073–1181.

Shen, T.; Jin, R.; Huang, Y.; Liu, C.; Dong, W.; Guo, Z.; Wu, X.; Liu, Y.; and Xiong, D. 2023. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*.

Song, F.; Yu, B.; Li, M.; Yu, H.; Huang, F.; Li, Y.; and Wang, H. 2024. Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18990–18998.

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

Sun, H.; Shen, Y.; and Ton, J.-F. 2024. Rethinking bradley-terry models in preference-based reward modeling: Foundations, theory, and alternatives. *arXiv preprint arXiv:2411.04991*.

Uma, A. N.; Fornaciari, T.; Hovy, D.; Paun, S.; Plank, B.; and Poesio, M. 2022. Learning from Disagreement: A Survey. *J. Artif. Int. Res.*, 72: 1385–1470.

van der Meer, M.; Falk, N.; Murukannaiah, P. K.; and Liscio, E. 2024. Annotator-Centric Active Learning for Subjective NLP Tasks. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 18537–18555. Miami, Florida, USA: Association for Computational Linguistics.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023a. Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv:2203.11171.

Wang, Y.; Zhong, W.; Li, L.; Mi, F.; Zeng, X.; Huang, W.; Shang, L.; Jiang, X.; and Liu, Q. 2023b. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Wang, Z.; Dong, Y.; Delalleau, O.; Zeng, J.; Shen, G.; Egert, D.; Zhang, J.; Sreedhar, M. N.; and Kuchaiev, O. 2024. Helpsteer 2: Open-source dataset for training top-performing reward models. *Advances in Neural Information Processing Systems*, 37: 1474–1501.

Warrens, M. J. 2012. Cohen's quadratically weighted kappa is higher than linearly weighted kappa for tridiagonal agreement tables. *Statistical Methodology*, 9(3): 440–444.

Xie, Z.; Wu, J.; Shen, Y.; Xia, Y.; Li, X.; Chang, A.; Rossi, R.; Kumar, S.; Majumder, B. P.; Shang, J.; et al. 2025. A survey on personalized and pluralistic preference alignment in large language models. *arXiv preprint arXiv:2504.07070*.

Yager, R. R. 1993. Families of OWA operators. *Fuzzy sets and systems*, 59(2): 125–148.

Zhang, M. J.; Wang, Z.; Hwang, J. D.; Dong, Y.; Delalleau, O.; Choi, Y.; Choi, E.; Ren, X.; and Pyatkin, V. 2024. Diverging Preferences: When do Annotators Disagree and do Models Know? *arXiv preprint arXiv:2410.14632*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023a. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Red Hook, NY, USA: Curran Associates Inc.

Zheng, R.; Dou, S.; Gao, S.; Hua, Y.; Shen, W.; Wang, B.; Liu, Y.; Jin, S.; Liu, Q.; Zhou, Y.; et al. 2023b. Secrets of rlhf in large language models part i: Ppo. *arXiv preprint arXiv:2307.04964*.

Zollo, T. P.; Siah, A. W. T.; Ye, N.; Li, A.; and Namkoong, H. 2024. Personalllm: Tailoring llms to individual preferences. *arXiv preprint arXiv:2409.20296*.