Operationalizing Pluralistic Values in Large Language Model Alignment Reveals Trade-offs in Safety, Inclusivity, and Model Behavior

Dalia Ali[1], Dora Zhao[2], Allison Koenecke[3], Orestis Papakyriakopoulos[1]
[1]Technical University of Munich, Germany
[2]Stanford University, USA
[3]Cornell University, USA
dalia.ali@tum.de, dorothyz@stanford.edu, koenecke@cornell.edu, orestis.p@tum.de

Abstract

Although large language models (LLMs) are increasingly trained using human feedback for safety and alignment with human values, alignment decisions often overlook human social diversity. This study examines how incorporating pluralistic values affects LLM behavior by systematically evaluating demographic variation and design parameters in the alignment pipeline. We collect alignment data from US and German participants *(N = 1,095 participants, 27,375 ratings)* who rated LLM responses across five dimensions: *Toxicity, Emotional Awareness (EA), Sensitivity, Stereotypical Bias, and Helpfulness.* We fine-tuned multiple Large Language Models and Large Reasoning Models using preferences from different social groups while varying rating scales, disagreement handling methods, and optimization techniques. The results revealed systematic demographic effects: male participants rated responses 18% less toxic than female participants; conservative and Black participants rated responses 27.9% and 44% higher on EA than liberal and White participants, respectively. Models fine-tuned on group-specific preferences exhibited distinct behaviors. Technical design choices showed strong effects: the preservation of rater disagreement achieved roughly 53% greater toxicity reduction than majority voting, and 5-point scales yielded about 22% more reduction than binary formats; and, Direct Preference Optimization (DPO) consistently outperformed Group Relative Policy Optimization (GRPO) in multi-value optimization. These findings represent a preliminary step in answering a critical question: *How should alignment balance expert-driven and user-driven signals to ensure both safety and fair representation?*

Extended version —https://arxiv.org/pdf/2511.14476

Introduction

As Large Language Models (LLMs) are deployed across real-world applications, aligning them with human values has become a core technical and ethical challenge (Wang et al. 2023; Liu et al. 2024b; Ouyang et al. 2022). Yet human values are not monolithic; they reflect diverse and often conflicting beliefs shaped by culture, politics, and lived experience (Chen, Zahidi, and Lespinet-Najib 2022; Khamassi, Nahon, and Chatila 2024; Hadar-Shoval et al. 2024). It is no longer realistic to assume that a single alignment objective can represent everyone (Gabriel and Keeling 2025; Sorensen et al. 2024), particularly given concerns about whose voices shape AI safety research (Lazar and Nelson 2023). Despite the recognition of value pluralism, alignment norms are often defined by small groups of developers, which risks excluding underrepresented worldviews (Kirk et al. 2024a). This is particularly evident in methods such as Constitutional AI, which bypass human disagreement by specifying fixed normative principles in advance (Bai et al. 2022b). While such approaches aim for consistency, they risk enforcing narrow value sets that overlook alternative perspectives.

As Gabriel (2020) argues, the central challenge is not simply deciding **what** values AI should align with, but identifying fair processes for deciding **whose** values matter in pluralistic societies. Recent work has begun to address this theoretically: Kasirzadeh (2024) distinguishes between first-order choices (how values like fairness are defined) and second-order questions (who defines them), while Sorensen et al. (2024) proposes formal strategies for integrating pluralism through group-specific fine-tuning.

However, a key empirical gap remains; recent studies show that current LLMs display far less preference variation than humans across cultural and political lines (Zhang et al. 2025), reinforcing an "algorithmic monoculture" that overlooks human value diversity (Kleinberg and Raghavan 2021). Modeling individual annotators, by contrast, helps recover minority viewpoints lost under majority voting (Gordon et al. 2022). Yet no work has examined how demographic diversity in feedback interacts with technical design choices to shape alignment outcomes. Building on advances in demographic and cultural alignment, we move from analysis to intervention.

Prior studies have compared GPT-4's safety annotations with human ratings across groups (Movva, Koh, and Pierson 2024), investigated cultural alignment via cross-lingual prompting and data mixtures (AlKhamissi et al. 2024), and introduced DICES (Aroyo et al. 2023), which frames disagreement as a signal for safety evaluation. Extending this work, we examine how demographic differences in feedback and design choices shape the collapse of pluralistic human values into a single model behavior during fine-tuning. Rather than implementing pluralistic alignment systems as proposed by Sorensen et al. (2024), we study how standard training homogenizes value variation and determines

which group's preferences dominate. **Specifically, we ask:** (1) How do models behave when aligned using feedback from different social groups? (2) How do technical choices such as rating scales, disagreement aggregation, and optimization methods affect learned values?

Our Contributions. We present a systematic empirical study of LLM alignment that jointly varies demographic composition and technical design using real human feedback (*27,375 ratings from 1,095 participants*). Models fine-tuned on feedback from Liberal, White, and Female participants show improvements of 5.0, 4.7, and 3.4 percentage points (relative to Conservative, Black, and Male baselines), across emotional awareness and toxicity. Technical choices yield even stronger effects: preserving disagreement improves toxicity reduction by 53% relative to majority voting, 5-point scales outperform binary formats by 22%, and DPO outperforms GRPO by about $8\times$ on toxicity and $3\times$ on emotional awareness. Together, these results demonstrate how demographic and design parameters jointly determine alignment behavior, advancing a framework for technically robust and socially inclusive alignment.

Related Work

Value Pluralism in AI Alignment Value pluralism presents a fundamental challenge for AI alignment. It holds that multiple and potentially conflicting moral values can each be valid without a single universal hierarchy, rejecting the idea of a final, universally agreed-upon solution to moral questions (Berlin 1969). This has prompted researchers to question alignment strategies built on assumptions of universal consensus. Sorensen et al. (2024) propose technical strategies to support pluralistic alignment, including steerable models that can be conditioned on specific perspectives at inference time, approaches for matching models to target population distributions, and methods for generating multiple reasonable responses. However, these frameworks remain largely theoretical, leaving open questions about how they perform when applied with real-world data and design constraints. Our work addresses this gap by empirically testing how demographic-specific feedback, rating formats, disagreement strategies, and optimization methods influence which value perspectives are amplified, suppressed, or erased in model behavior. This provides the first empirical grounding of value pluralism collapse in applied alignment design.

Limitations of Current Human Feedback Pipelines for Alignment Recent alignment pipelines have enhanced model safety by incorporating human preferences through norms such as helpfulness, honesty, and harmlessness (HHH) (Askell et al. 2021; Bai et al. 2022a; Ouyang et al. 2022). Datasets such as BeaverTails (Jin et al. 2023), PRISM (Kirk et al. 2024b), and OASST1 (Köpf et al. 2024) utilize crowd-sourced feedback to capture alignment signals, whereas benchmarks like BBQ (Parrish et al. 2021) and pipelines like GenderAlign (Zhang et al. 2024) focus on addressing fairness and stereotyping.

Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al. 2022) trains a reward model from human preference comparisons and uses it to optimize the model via reinforcement learning. However, most RLHF paradigms assume homogeneous preferences encodable by a single reward model (Park et al. 2024). This leads to implicit averaging that prioritizes majority preferences while neglecting minorities (Chakraborty et al. 2024). In extreme cases, this leads to preference collapse, where minority preferences are disregarded (Xiao et al. 2024). These pipelines overlook how different social groups interpret alignment concepts like harm or respect based on cultural or political context (Kovač et al. 2023; Sorensen et al. 2024; Pang et al. 2023; Lyu et al. 2025; Pan et al. 2025).

Recent findings on algorithmic monoculture highlight challenges in capturing human value diversity. Zhang et al. (2025) show that 21 state-of-the-art LLMs produce significantly less preference variation than humans across five countries *(N=15,000)*, limiting the diversity expressible in current preference datasets. While their work proposes sampling methods to diversify model outputs, our study addresses the complementary problem of how evaluation design, including demographic variation, rating scale format, and disagreement handling, can preserve or suppress pluralistic values within constrained response spaces.

Rating Scales and Disagreement Handling The choice of rating scale, whether Likert (e.g., a 5-point scale from "strongly disagree" to "strongly agree"), binary (e.g., "agree" or "disagree"), or pairwise (e.g., choosing a preferred response between two options), influences both rater behavior and model outcomes. Survey research shows that scale format affects response biases, including net acceptance, extreme responding, and misinterpretation of reverse-coded items (Weijters, Cabooter, and Schillewaert 2010). Likert scales are especially prone to central tendency bias, where respondents tend to avoid the endpoints and favor the midpoints (Douven and Schupbach 2018). In LLM alignment, recent studies increasingly use pairwise comparisons for training and evaluation (Zheng et al. 2023). Despite the widespread use of various formats in human feedback pipelines, little is known about how these choices impact model behavior. Our work fills this gap through a systematic comparison of rating formats.

Disagreement between raters is often resolved using majority voting or averaging, which can erase minority perspectives and obscure subjective variation (Gordon et al. 2022; Davani, Díaz, and Prabhakaran 2022). Recent work proposes alternatives such as soft-label representations that retain disagreement distributions, and multi-annotator models that predict individual judgments (Davani, Díaz, and Prabhakaran 2022; Aroyo and Welty 2015). Kraus and Kroll (2025) offers a framework for distinguishing noise from meaningful signal, arguing that tasks involving personal values should preserve disagreement rather than aggregate it. Our work builds on this literature by systematically comparing how different aggregation methods, from consensus-based to disagreement-preserving, shape the downstream behavior of aligned models.

Data Collection

Prompt Selection and Response Generation To examine how social group differences and technical choices affect model alignment, we develop a bilingual alignment pipeline (English-German) with a focus on gender-related scenarios. Crucially, all participants, regardless of their own demographics, view the same set of gender-related prompt-response pairs, isolating the effects of rater demographics and design choices without confounding topic variation. The prompts are drawn from red-teaming, gender bias, and alignment (BeaverTails) benchmarks (Ganguli et al. 2022; Parrish et al. 2021; Ji et al. 2023), using gender-related keyword filters. Responses are generated using `Wizard-Vicuna-7B-Uncensored-GPTQ` (TheBloke 2023), selected because (1) its lack of safety filtering ensures exposure to a broad range of potentially problematic outputs, and (2) it is a stable open-source base model ensuring reproducibility. This yields 37,884 prompt-response pairs, from which we select 1,761 unique pairs for human evaluation. The pairs are translated into German using DeepL (2024) and checked for semantic equivalence by a native speaker (see Supplementary A.1 and A.2).

Rating Dimensions and Scale Design Participants rate model outputs along five alignment dimensions: *Toxicity*, *Emotional Awareness*, *Sensitivity & Openness*, *Helpfulness*, and *Stereotypical Bias* (see Supplementary A.3 for definitions provided to participants). Following recent sociotechnical alignment frameworks (Kirk et al. 2023), these dimensions capture how AI responses affect users, balancing ethical concerns (toxicity and stereotypical bias) with social dimensions (sensitivity, openness, and helpfulness) (Liu et al. 2024a; Bilquise, Ibrahim, and Shaalan 2022; Yin, Goh, and Hu 2024; Lissak et al. 2024; Ji et al. 2023). This multidimensional approach expands beyond traditional helpfulness and harmfulness prompt–response pairs, yielding a total of 27,375 evaluations. Each dimension addresses aspects of human-AI interaction that different social groups may prioritize differently, making them relevant for studying value pluralism in alignment. Each dimension is rated using a 5-point Likert scale (from "Strongly Disagree" to "Strongly Agree"), allowing participants to express the intensity of their judgments rather than making binary classifications. Participants received clear definitions for each dimension (see supplementary A.3) before rating, ensuring consistent interpretation across cultural contexts.

Participant Recruitment and Demographics We obtain ethics approval from *Technical University of Munich* review board prior to data collection. We recruit 1,095 participants from the United States and Germany via Prolific (2025), targeting balanced distribution across gender, political spectrum, age, ethnicity, and country of residence (see Supplementary Table 3 for participant demographics). Each participant rates five prompt–response pairs. Participants are compensated and informed that their responses will help train future LLMs. We include attention checks to ensure data quality and collect demographic details (gender, age, ethnicity, political orientation) to support subgroup analysis (Supplementary A.4).

Data Analyses Informing Experiment

Pervasive Disagreement Across All Dimensions Analysis of unique 1,761 prompt–response pairs evaluated by multiple participants *(mean: 3.1 raters per pair, all pairs with $\geq$2 raters)* reveals systematic disagreement across all alignment dimensions. Disagreement, defined as variation in 5-point Likert ratings within the same prompt–response pair, occurred in 85.3% of cases, ranging from 84.5% for sensitivity to 86.2% for helpfulness. This analysis of 5,475 individual ratings from 1,095 participants demonstrates that conflicting human preferences are fundamental characteristics of alignment evaluation, not isolated incidents.

Cross-Dimensional Rating Complexity Participants frequently assign seemingly contradictory ratings to the same response across different alignment dimensions. By examining how often responses rated positively for one dimension were also rated positively for another, we found that among responses rated as emotionally aware, 20.8% were also rated as toxic, and 33.5% were also rated as stereotypically biased. This pattern shows that participants can simultaneously perceive both positive and negative qualities in the same AI response, suggesting that alignment evaluation involves complex trade-offs rather than straightforward categorization (Figure 1). This aligns with evidence that users hold diverse, sometimes conflicting preferences rather than a single aggregated signal (Gölz, Haghtalab, and Yang 2025; Fleisig, Fazelpour et al. 2025).

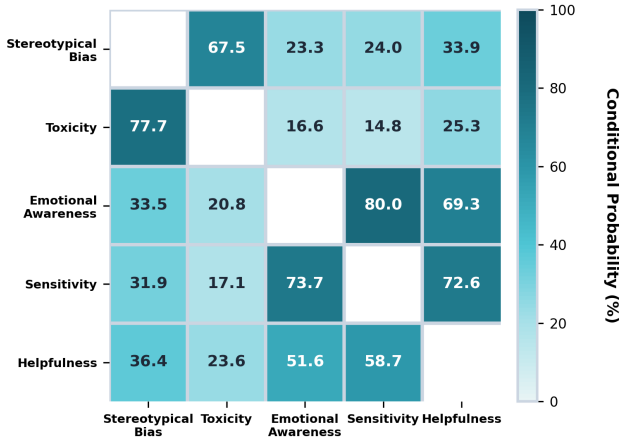

Figure 1: Conditional probabilities between alignment dimensions. Each cell shows the probability of a positive rating in the column given a positive rating in the row.

Ordinal Regression Results

We conduct a cumulative link mixed model (CLMM) analysis (Christensen 2023) to identify demographic factors that significantly influence alignment ratings across the five di-

mensions. Our model specification is:

$$CLMM(AlignmentRating_i \sim Country_i + Gender_i + Age_i + PoliticalSpectrum_i + Ethnicity_i + (1 \mid ParticipantID_i) + (1 \mid Context_i)) \quad (1)$$

The reference group is defined as *White, US-based, Liberal, aged 18–30, and identifying as she/her/hers*. Our analysis reveals several statistically significant demographic effects (see Table **??**). Male participants rate responses as 18% less toxic and 20.9% less stereotypically biased compared to female participants. Conservative participants perceive responses as 27.9% more sensitive and 27% more emotionally aware than Liberals. Black or African American participants rate responses as 58% more sensitive and 44% more emotionally aware than White participants, a pattern consistent with work showing that Black Americans often attend more closely to emotional nuance and cultural sensitivity in AI interactions (Sandoval et al. 2025; Basoah et al. 2025). Additionally, participants aged 51-60 find responses 40.6% less helpful compared to younger participants (see Supplementary A.6).

Dimension	Gender	Age	Political	Ethnicity
Toxicity	**Yes**** (M ↓)	No	No	No
Helpfulness	No	**Yes***** (51-60 ↓)	No	No
Sensitivity	No	No	**Yes**** (C ↑)	**Yes***** (B ↑)
Stereotypical Bias	**Yes***** (M ↓)	No	No	No
Emotional Awareness	No	No	**Yes*** (C ↑)	**Yes***** (B ↑)

Table 1: Demographic Predictors of Participant Ratings. Arrows show direction relative to baseline (↑ higher, ↓ lower); *M = Male, C = Conservative, B = Black or African American* (baselines: White, US-based, Liberal, aged 18–30, and female). *p $<$ .05, **p $<$ .01, ***p $<$ .001.

Experiment Setup

Fine-tuning Experiments

We conduct four fine-tuning experiments to examine how demographic variation and technical design choices affect model alignment. Experiments 1–3 rely on Direct Preference Optimization (DPO) (Rafailov et al. 2023), which learns from pairwise response comparisons. We use DPO for Experiments 1–3 because Experiment 4 shows it outperforms GRPO. Experiment 4 then contrasts DPO with Group-Relative Policy Optimization (GRPO) (Shao et al. 2024), an on-policy method that optimizes scalar rewards across multiple sampled completions using group-normalized advantages. Each method requires distinct data formatting, and we construct datasets accordingly. All experiments focus on toxicity and emotional awareness; these two dimensions are chosen to capture complementary safety concerns of harm avoidance and social understanding, where demographic and cultural backgrounds shape perceptions. Across the dataset, we obtain 5,475 ratings per dimension from 1,095 raters, applied to 1,761 unique prompt–response pairs, some of which receive multiple independent ratings (see Supplementary B.1, B.2 and B.3). All experiments are replicated across seven model architectures to ensure robustness (see Supplementary Table 9 for model overviews).

Experiment 1: Data Stratification by Demographic Groups We fine-tune models on balanced subsets of human feedback to examine how demographic composition influences alignment outcomes. Three contrasts are tested: gender (female vs. male), political orientation (liberal vs. conservative), and ethnicity (White vs. Black).[1] Each subgroup contributed an equal number of ratings to control for the size of the data set and distributional effects. Inter-annotator reliability, measured using Krippendorff's α for toxicity annotations, was α=0.35 (White) and α=0.44 (Black) for ethnicity, α=0.39 (Liberal) and α=0.33 (Conservative) for political orientation, and α=0.38 for both female and male raters, consistent with prior work (Ross et al. 2016; Bui, von der Wense, and Lauscher 2025); these values indicate substantial within-group variation rather than consensus. For each subgroup, a separate model is fine-tuned using DPO while preserving all five-point Likert ratings. The resulting models are evaluated on two alignment dimensions: Toxicity and Emotional Awareness, and compared against a baseline trained on feedback from conservative, male, and Black participants. Following Table **??**, we focus on emotional awareness for the political spectrum and ethnicity, and on toxicity for gender, with complementary robustness checks (Supplementary B.6). This design enables systematic analysis of both within and cross-dimension effects.

Experiment 2: Rating Scale Granularity We investigate how rating scale granularity influences alignment outcomes by creating three versions of the toxicity dataset: (1) a 5-point Likert scale (Strongly Disagree to Strongly Agree), (2) a 3-point scale (Disagree, Neutral, Agree), and (3) a binary scale (excluding Neutral responses). For each scale version, we retain all participant ratings and fine-tune separate models using DPO on toxicity feedback. This setup enables us to assess how the level of granularity impacts the model's ability to learn alignment preferences. We focus on Likert scale ratings rather than pairwise comparisons to capture the full spectrum of participant sentiment and enable analysis across multiple granularity levels.

Experiment 3: Disagreement Handling Strategies To examine the effects of inter-annotator disagreement handling, we create five training datasets using different aggregation methods: (1) all data without aggregation, (2) majority vote, (3) full consensus (complete rater agreement only), (4) random selection (first rater's rating), and (5) averaged ratings (mean of numeric-coded responses rounded to nearest category). We fine-tune separate DPO models on each dataset to assess how disagreement resolution strategies impact alignment performance. All models are trained using the 5-point Likert scale and evaluated on the Toxicity dimension.

[1] Age affects helpfulness only; it shows no consistent effects for toxicity or emotional awareness, so we do not use it as a contrast.

Experiment 4: DPO vs. GRPO Optimization Method Comparison We systematically compare DPO and GRPO in multi-value optimization using our combined Toxicity and Emotional Awareness dataset (5-point scale, all data without aggregation). We format the merged dataset according to each method's requirements and fine-tune separate models to evaluate their comparative effectiveness in simultaneously reducing toxicity and improving emotional awareness in model outputs.

Experimental Design and Statistical Analysis All experiments use LoRA fine-tuning (Hu et al. 2022), with deterministic sampling (temperature=0.0) for reproducible evaluation across seven diverse model architectures (1B-14B parameters), ensuring consistent comparison of alignment effects across demographic groups and technical conditions. After fine-tuning the models on human feedback, we use GPT–4o–mini to score their toxicity and emotional-awareness outputs. We validate these LLM-generated scores against human expert judgements, achieving 85% agreement (see Supplementary B.2 and B.3). We use a DerSimonian–Laird random-effects meta-analysis (DerSimonian and Laird 1986) to pool effects across model architectures, accounting for both within-model variance and between-model heterogeneity. Full equations and derivations are provided in (Supplementary B.5).

Results

Experiment 1: Demographic Composition Effects on Model Fine-tuning Outcomes

We assess how subgroup-specific fine-tuning affects alignment outcomes across gender, political orientation, and ethnicity by comparing models trained on *female, liberal, and White* feedback with models trained on *male, conservative, and Black* feedback (Figure 2).

Demographic Feedback Shapes Specific Behaviors. Models fine-tuned on Liberal and White feedback produced higher emotional awareness scores than those trained on Conservative and Black feedback (pooled effects: 0.049, $p = 0.010$, and 0.046, $p = 0.001$, respectively). Similarly, models fine-tuned on Female feedback showed lower toxicity than those trained on Male feedback (pooled effect: -0.035, $p = 0.002$). These effects were consistent across seven model architectures.

Effects Are Dimension-Specific. To test for generalization beyond the target dimensions, each demographically fine-tuned model was evaluated on both toxicity and emotional awareness. No statistically significant cross-dimensional effects were observed. For instance, the model fine-tuned on Female toxicity feedback did not significantly alter emotional awareness ($p = 0.860$), while Liberal and White emotional-awareness models showed no effect on toxicity ($p = 0.500$ and $p = 0.880$, respectively). These findings demonstrate that demographic composition produces measurable yet dimension-specific effects on alignment: subgroup value preferences are reliably encoded in fine-tuned models without introducing unintended behavioral shifts across unrelated alignment dimensions.

Experiment 2: Rating Scale Granularity Effects on Alignment Training

We examine how the granularity of the rating scale affects alignment effectiveness by comparing 5-point, 3-point, and binary scales using data from the toxicity dimension and DPO fine-tuning. We examine scale granularity with a focus on toxicity ratings (Figure 3).

Granular Scales Improve Alignment Performance All scales produce a reduction in toxicity relative to the control model (no fine-tuning), but with substantially different effect sizes. The 5-point scale achieves the largest effect (-0.242), followed by the 3-point scale (-0.225) and binary scale (-0.198). Pairwise comparisons reveal that the 5-point scale outperforms the binary scale ($p = 0.0141$) and marginally outperforms the 3-point scale ($p = 0.0140$), indicating that scale granularity has an impact on model training outcomes. The difference between 3-point and binary scales is statistically marginal yet directionally consistent, confirming that finer scales yield stronger alignment effects.

Implications of Scale Granularity for Model Learning 5-point scales are 22% more effective than binary scales at reducing toxicity (effect sizes: -0.242 vs. -0.198); this pattern is consistent with findings that multi-level cardinal feedback (such as 5-point scale) yields more learnable reward models than pairwise preferences, while maintaining similar levels of inter-rater reliability (Kreutzer, Uyheng, and Riezler 2018). This performance gap reveals that reducing human feedback to binary choices for alignment tasks underutilizes available preference information. Although binary formats are widely used, our findings show measurable costs to the effectiveness of alignment.

Experiment 3: Disagreement Handling Strategy Effects on Alignment Training

We examine how approaches to handling inter-annotator disagreement affect alignment outcomes by comparing five aggregation strategies: preserving all ratings, averaging, majority vote, random selection, and full consensus (Figure 4).

Complete Rating Preservation Demonstrates Superior Alignment Performance All strategies reduce toxicity relative to the control model, but their effectiveness varies substantially. Preserving all ratings yields the strongest reduction (-0.242), closely followed by averaging ratings (-0.229). Other approaches perform less effectively: majority vote (-0.158), random selection (-0.146), and full consensus (-0.039). Pairwise comparisons confirm that preserving all ratings consistently outperforms consensus-based and random selection methods, while the averaged-rating strategy performs comparably to full preservation. Preserving all ratings is approximately 53% more effective than majority vote and nearly 6× more effective than full consensus in reducing toxicity. The strong performance of preserving or averaging all ratings shows that disagreement carries meaningful signal. Majority vote and consensus filtering suppress minority perspectives and weaken alignment. Treating disagreement as information rather than

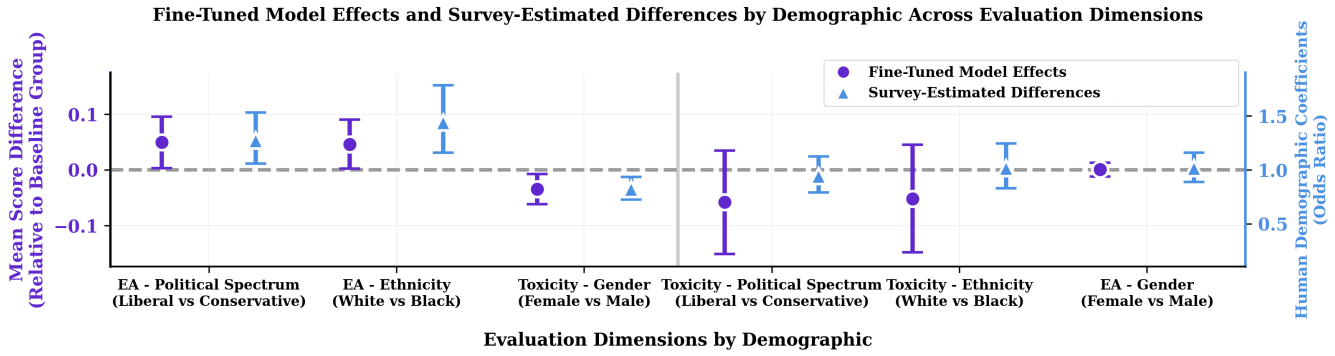

Figure 2: Effects of demographic composition on fine-tuning outcomes. Purple circles show model differences between groups (Liberal, Female, White minus Conservative, Male, Black, respectively) on Emotional Awareness or Toxicity ratings; blue triangles show survey-estimated differences for the same contrasts. Positive values indicate higher Emotional Awareness for the first group; negative values indicate lower Toxicity for the first group. Error bars show 95% confidence intervals.

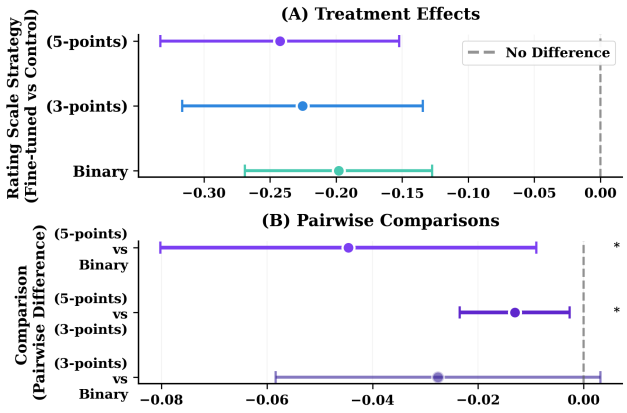

Figure 3: Rating Scale Effects on Toxicity Reduction. (A) All scales reduce toxicity relative to the control (no fine-tuning), with the 5-point most effective. (B) 5-point scales significantly outperform binary. Error bars: 95% CIs. Lower values indicate reduced toxicity.

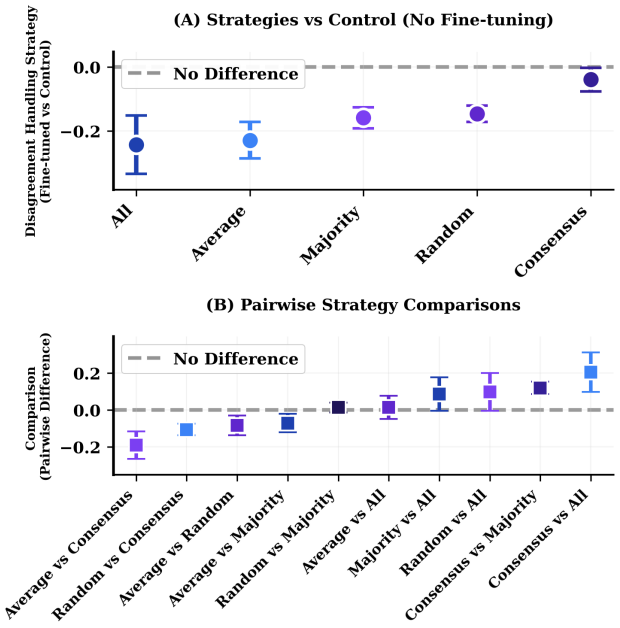

Figure 4: Disagreement Handling Strategy Effects on Alignment Training. (A) Strategy performance relative to the control (no fine-tuning) on toxicity. (B) Pairwise strategy comparisons. Error bars show 95% CIs. Lower values indicate better performance.

noise improves robustness and inclusivity, yielding stronger alignment across diverse preferences.

Experiment 4: DPO vs GRPO Comparison

DPO outperforms GRPO Training We compare DPO and GRPO using a multi-objective loss that jointly optimizes toxicity and emotional awareness (Figure 5). When trained on the same-sized dataset, DPO consistently outperforms GRPO across both dimensions. The DPO-Toxic+EA model achieves toxicity reduction (-0.159, $p < 0.001$) and emotional awareness improvement (0.084), while GRPO-Toxic+EA shows only marginal effects (-0.020, $p = 0.045$; 0.029, $p = 0.034$). These correspond to effect sizes roughly $8\times$ larger for toxicity and nearly $3\times$ larger for emotional awareness, demonstrating DPO's advantage over GRPO for multivalue alignment. All effects relative to the control (no fine-tuning) are statistically significant ($p < 0.05$).

Single-Objective DPO Outperforms Multi-Objective Training Comparing DPO variants reveals clear trade-offs between single- and multi-objective training. For toxicity reduction, DPO-Toxic yields the largest effect (-0.243, $p < 0.001$), significantly outperforming both DPO-EA (difference: 0.097, $p < 0.001$) and the multi-objective DPO-Toxic+EA model (difference: -0.084, $p = 0.022$). DPO-EA and DPO-Toxic+EA exhibit comparable toxicity effects (difference: 0.013, $p = 0.743$). For emotional awareness, all three DPO variants perform similarly: DPO-EA achieves

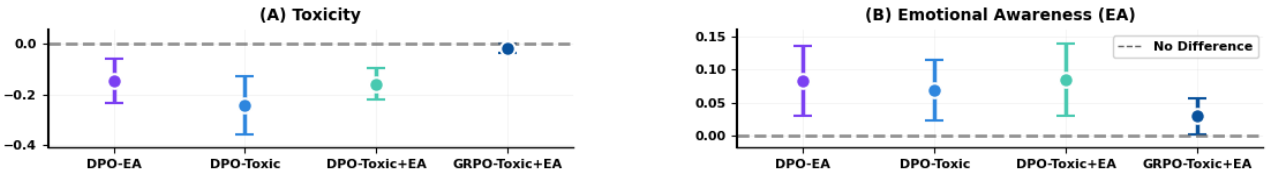

Figure 5: DPO and GRPO Optimization Methods Comparison. (A-B) Performance of each DPO and GRPO trained model. (C-D) Pairwise comparisons between single-objective and multi-objective DPO approaches. Error bars show 95% CIs.

0.083 ($p = 0.002$), DPO-Toxic achieves 0.068 ($p = 0.003$), and DPO-Toxic+EA achieves 0.084 ($p = 0.003$), with no significant pairwise differences (all $p > 0.13$). These findings show that single-objective toxicity fine-tuning maximizes performance on its targeted dimension, whereas emotional-awareness gains remain indistinguishable across all models optimized with DPO.

Optimization Method and Training Objective Matter Two findings emerge from our dataset. First, the choice of optimization method significantly influences alignment outcomes: DPO yields larger and more reliable gains than GRPO on the same data. Second, focused single-objective DPO fine-tuning produces stronger and more interpretable effects than multi-objective training. Taken together, these results challenge the assumption that multi-objective alignment or newer optimization methods automatically perform better, highlighting the need for methodological precision in preference-based fine-tuning.

Discussion and Conclusion

How should alignment processes balance expert-driven and user-driven signals to ensure both safety and fair representation? Lazar and Nelson (2023) argue that AI safety is shaped by a demographic monoculture that lacks legitimacy and intellectual breadth, while Gyevnar and Kasirzadeh (2025) call for an epistemically inclusive and pluralistic approach. Anthis et al. (2025) show that even a single rigorous fairness criterion becomes intractable for general-purpose LLMs across diverse contexts, and Kleinberg, Mullainathan, and Raghavan (2016) formally demonstrate that core group-fairness conditions cannot be satisfied simultaneously except in highly restricted cases. Taken together with our empirical results, this suggests that known pathologies in algorithmic fairness also appear in alignment research, with limited participation and structural limits jointly produce systematic distortions in model behavior.

Safety judgments are not universal but reflect specific demographic perspectives. Male and female participants rated identical responses with an 18% difference in perceived toxicity, while Conservative and Black participants reported 27.9% and 44% higher emotional awareness ratings, respectively, compared to Liberal, White participants. Current alignment approaches that aggregate these differences away may systematically exclude safety-relevant perspectives. Preserving disagreement achieved the strongest toxicity reduction, outperforming the majority vote by 53%, and other strategies. This suggests alignment annotators' "noise" may be an essential safety signal. Technical choices such as rating scales (5-point exceeding binary by 22%) and optimization methods (DPO exceeding GRPO) profoundly impact safety performance. Rather than trading off safety against inclusivity, we find that inclusive approaches enhance safety outcomes.

Our findings reveal systematic demographic effects: models trained on White, Liberal, and Female feedback achieve higher emotional awareness and lower toxicity respectively than those trained on Black, Conservative, and Male feedback. These shifts occur because demographic groups differ fundamentally in how they evaluate harm and emotional quality. Together, these results show that demographic diversity is not a one-time dataset choice but an ongoing alignment requirement. Safety judgments vary across populations and shift over time, demanding continuous reassessment of *whose* perspectives are prioritized. Since both training data and technical design systematically advantage certain groups over others, robust alignment requires periodic audits: *Which demographic groups dominate our data? How do our methodological decisions suppress minority voices?* This reflexive practice can help ensure that alignment does not unintentionally center the values of specific groups.

Rather than relying only on prescriptive value choices made by researchers or model developers (Kirk et al. 2024a), expert technical knowledge should serve democratic inclusion rather than replace it. Our findings support systems that preserve diverse safety perspectives rather than require complete agreement. Robust AI safety requires both expert sophistication and comprehensive democratic representation as complementary requirements for effective alignment.

We conclude with limitations (see more in Supplementary A.5). Our dataset covers only two WEIRD-dominant countries and has uneven demographic representation (underrepresenting conservatives, gender minorities, and older adults). Model evaluation used GPT-4o-mini, though its judgments are in high agreement with human reviewers. Our optimization analysis focused on DPO and GRPO, leaving other methods for future work. These constraints do not affect the core patterns but highlight the need for more diverse datasets and broader evaluation of optimization approaches.

Acknowledgements

We thank Michèle Wieland, Aysenur Kocak, Cheng Yu, Furkan Kadıoğlu, Jana Diesner, Shaghayegh Kolli and Nafiseh Nikeghbal for their valuable discussions and support throughout this project.

References

AlKhamissi, B.; ElNokrashy, M.; AlKhamissi, M.; and Diab, M. 2024. Investigating cultural alignment of large language models. *arXiv preprint arXiv:2402.13231*.

Anthis, J. R.; Lum, K.; Ekstrand, M.; Feller, A.; and Tan, C. 2025. The impossibility of fair LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 105–120.

Aroyo, L.; Taylor, A.; Diaz, M.; Homan, C.; Parrish, A.; Serapio-García, G.; Prabhakaran, V.; and Wang, D. 2023. Dices dataset: Diversity in conversational ai evaluation for safety. *Advances in Neural Information Processing Systems*, 36: 53330–53342.

Aroyo, L.; and Welty, C. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1): 15–24.

Askell, A.; Bai, Y.; Chen, A.; Drain, D.; Ganguli, D.; Henighan, T.; Jones, A.; Joseph, N.; Mann, B.; DasSarma, N.; Elhage, N.; Hatfield-Dodds, Z.; Hernandez, D.; Kernion, J.; Ndousse, K.; Olsson, C.; Amodei, D.; Brown, T.; Clark, J.; McCandlish, S.; Olah, C.; and Kaplan, J. 2021. A General Language Assistant as a Laboratory for Alignment. Accessed: 2025-05-29, arXiv:2112.00861.

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

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022b. Constitutional AI: harmlessness from AI feedback. 2022. *arXiv preprint arXiv:2212.08073*, 8(3).

Basoah, J.; Cunningham, J. L.; Adams, E.; Bose, A.; Jain, A.; Yadav, K.; Yang, Z.; Reinecke, K.; and Rosner, D. 2025. Should AI mimic people? Understanding AI-Supported writing technology among Black users. *Proceedings of the ACM on Human-Computer Interaction*, 9(7): 1–51.

Berlin, I. 1969. Two Concepts of Liberty. In *Four Essays on Liberty*, 118–172. Oxford University Press.

Bilquise, G.; Ibrahim, S.; and Shaalan, K. F. 2022. Emotionally Intelligent Chatbots: A Systematic Literature Review. *Human Behavior and Emerging Technologies*. Accessed: 2024-07-21.

Bui, M. D.; von der Wense, K.; and Lauscher, A. 2025. Multi3Hate: Multimodal, Multilingual, and Multicultural Hate Speech Detection with Vision–Language Models. *arXiv preprint arXiv:2411.03888*.

Chakraborty, S.; Qiu, J.; Yuan, H.; Koppel, A.; Manocha, D.; Huang, F.; Bedi, A. S.; and Wang, M. 2024. MaxMin-RLHF: Alignment with Diverse Human Preferences. *Proceedings of the 41st International Conference on Machine Learning*.

Chen, L.; Zahidi, K.; and Lespinet-Najib, V. 2022. Aligning artificial intelligence with human values: reflections from a phenomenological perspective. *AI & Society*, 37(3): 1151–1171.

Christensen, R. H. B. 2023. *ordinal—Regression Models for Ordinal Data*. R package version 2023.12-4.1.

Davani, A. M.; Díaz, M.; and Prabhakaran, V. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10: 92–110.

DeepL. 2024. DeepL Translator. Accessed: 2024-07-26.

DerSimonian, R.; and Laird, N. 1986. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3): 177–188.

Douven, I.; and Schupbach, J. N. 2018. A Bayesian perspective on Likert scales and central tendency. *Psychonomic Bulletin & Review*, 25(3): 1203–1211.

Fleisig, J.; Fazelpour, S.; et al. 2025. Perspectival Homogenization: A Normative Framework for Reasoning About Disagreement in AI Development. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

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

Gabriel, I.; and Keeling, G. 2025. A matter of principle? AI alignment as the fair treatment of claims. *Philosophical Studies*.

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

Gölz, P.; Haghtalab, N.; and Yang, K. 2025. Distortion of AI Alignment: Does Preference Optimization Optimize for Preferences? *arXiv preprint arXiv:2502.04564*.

Gordon, M. L.; Lam, M. S.; Park, J. S.; Patel, K.; Hancock, J. T.; Hashimoto, T.; and Bernstein, M. S. 2022. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems*, CHI '22, 1–19. New York, NY, USA: ACM. ISBN 978-1-4503-9157-3/22/04.

Gyevnar, B.; and Kasirzadeh, A. 2025. AI safety for everyone. *Nature Machine Intelligence*, 1–12.

Hadar-Shoval, D.; Asraf, K.; Mizrachi, Y.; Haber, Y.; and Elyoseph, Z. 2024. Assessing the Alignment of Large Language Models With Human Values for Mental Health Integration: Cross-Sectional Study Using Schwartz's Theory of Basic Values. *JMIR Mental Health*, 11: e55988.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Ji, J.; Liu, M.; Dai, J.; Pan, X.; Zhang, C.; Bian, C.; Chen, B.; Sun, R.; Wang, Y.; and Yang, Y. 2023. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36: 24678–24704.

Jin, D.; Mehri, S.; Hazarika, D.; Padmakumar, A.; Lee, S.; Liu, Y.; and Namazifar, M. 2023. Data-efficient alignment of large language models with human feedback through natural language. *arXiv preprint arXiv:2311.14543*.

Kasirzadeh, A. 2024. Plurality of value pluralism and ai value alignment. In *Pluralistic Alignment Workshop at NeurIPS 2024*.

Khamassi, M.; Nahon, M.; and Chatila, R. 2024. Strong and weak alignment of large language models with human values. *Scientific Reports*, 14(1): 19399.

Kirk, H.; Vidgen, B.; Röttger, P.; et al. 2024a. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 6: 383–392.

Kirk, H. R.; Vidgen, B.; Röttger, P.; and Hale, S. A. 2023. The Empty Signifier Problem: Towards Clearer Paradigms for Operationalising "Alignment" in Large Language Models. Accessed: 2024-08-21, arXiv:2310.02457.

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

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Kleinberg, J.; and Raghavan, M. 2021. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22): e2018340118.

Köpf, A.; Kilcher, Y.; von Rütte, D.; Anagnostidis, S.; Tam, Z. R.; Stevens, K.; Barhoum, A.; Nguyen, D.; Stanley, O.; Nagyfi, R.; et al. 2024. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36.

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

Kraus, K.; and Kroll, M. 2025. Maximizing Signal in Human-Model Preference Alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 27392–27400.

Kreutzer, J.; Uyheng, J.; and Riezler, S. 2018. Reliability and learnability of human bandit feedback for sequence-to-sequence reinforcement learning. *arXiv preprint arXiv:1805.10627*.

Lazar, S.; and Nelson, A. 2023. AI safety on whose terms?

Lissak, S.; Calderon, N.; Shenkman, G.; Ophir, Y.; Fruchter, E.; Klomek, A. B.; and Reichart, R. 2024. The Colorful Future of LLMs: Evaluating and Improving LLMs as Emotional Supporters for Queer Youth. Accessed: 2024-06-27, arXiv:2402.11886.

Liu, Y.; Yao, Y.; Ton, J.-F.; Zhang, X.; Guo, R.; Cheng, H.; Klochkov, Y.; Taufiq, M. F.; and Li, H. 2024a. Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment. Accessed: 2024-06-25, arXiv:2308.05374.

Liu, Y.; Zhou, H.; Guo, Z.; Shareghi, E.; Vulić, I.; Korhonen, A.; and Collier, N. 2024b. Aligning with Human Judgement: The Role of Pairwise Preference in Large Language Model Evaluators. In *Proceedings of the Conference on Language Modeling (COLM)*.

Lyu, H.; Luo, J.; Kang, J.; and Koenecke, A. 2025. Characterizing Bias: Benchmarking Large Language Models in Simplified versus Traditional Chinese. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 2815–2846.

Movva, R.; Koh, P. W.; and Pierson, E. 2024. Annotation alignment: Comparing LLM and human annotations of conversational safety. *arXiv preprint arXiv:2406.06369*.

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

Pan, E.; Choi, A. S. G.; Ter Hoeve, M.; Seto, S.; and Koenecke, A. 2025. Analyzing Dialectical Biases in LLMs for Knowledge and Reasoning Benchmarks. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Findings of the Association for Computational Linguistics: EMNLP 2025*, 20882–20893. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-335-7.

Pang, R. Y.; Cenatempo, J.; Graham, F.; Kuehn, B.; Whisenant, M.; Botchway, P.; Stone Perez, K.; and Koenecke, A. 2023. Auditing cross-cultural consistency of human-annotated labels for recommendation systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1531–1552.

Park, C.; Liu, M.; Zhang, K.; and Ozdaglar, A. 2024. Principled RLHF from Heterogeneous Feedback via Personalization and Preference Aggregation. *arXiv preprint arXiv:2405.00254*.

Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Htut, P. M.; and Bowman, S. R. 2021. BBQ: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193*.

Prolific. 2025. Prolific: Online participant recruitment for surveys and experiments. https://www.prolific.com/. Accessed: 12 November 2025.

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.

Ross, B.; Rist, M.; Carbonell, G.; Cabrera, B.; Kurowsky, N.; and Wojatzki, M. 2016. Measuring the reliability of hate speech annotations: The case of the European refugee crisis. *arXiv preprint arXiv:1701.08118*.

Sandoval, S. C.; Acquaye, C.; Cobbina, K.; Teli, M. N.; and Daumé III, H. 2025. My LLM might Mimic AAE–But When Should it? *arXiv preprint arXiv:2502.04564*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath:

Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

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*.

TheBloke. 2023. Wizard-Vicuna-7B-Uncensored-GPTQ. https://huggingface.co/TheBloke/Wizard-Vicuna-7B-Uncensored-GPTQ. Accessed July 2024.

Wang, Y.; Zhong, W.; Li, L.; Mi, F.; Zeng, X.; Huang, W.; Shang, L.; Jiang, X.; and Liu, Q. 2023. Aligning Large Language Models with Human: A Survey. *arXiv preprint arXiv:2307.12966*.

Weijters, B.; Cabooter, E.; and Schillewaert, N. 2010. The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, 27(3): 236–247.

Xiao, J.; Li, Z.; Xu, M.; Zhang, C.; Wang, M.; and Liu, Z. 2024. On the Algorithmic Bias of Aligning Large Language Models with RLHF: Preference Collapse and Matching Regularization. *arXiv preprint arXiv:2405.16455*.

Yin, J.; Goh, T.-T.; and Hu, Y. 2024. Interactions with educational chatbots: the impact of induced emotions and students' learning motivation. *International Journal of Educational Technology in Higher Education*, 21(1): 47.

Zhang, L. H.; Milli, S.; Jusko, K.; Smith, J.; Amos, B.; Bouaziz, W.; Revel, M.; Kussman, J.; Titus, L.; Radharapu, B.; Yu, J.; Sarma, V.; Rose, K.; and Nickel, M. 2025. Cultivating Pluralism In Algorithmic Monoculture: The Community Alignment Dataset. *arXiv preprint arXiv:2507.09650*.

Zhang, T.; Zeng, Z.; Xiao, Y.; Zhuang, H.; Chen, C.; Foulds, J.; and Pan, S. 2024. GenderAlign: An Alignment Dataset for Mitigating Gender Bias in Large Language Models. *arXiv preprint arXiv:2406.13925*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623.