Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Special Track on Human-Centred AI

DiverValue-Bench: A Benchmark and Fine-Tuning Framework for Aligning Large Language Models with Diverse Human Values

Yao Liang[1,5], **Dongcheng Zhao**[1], **Feifei Zhao**[1,2,5*], **Guobin Shen**[1,6], **Yuwei Wang**[1], **Dongqi Liang**[1,5] and **Yi Zeng**[1,2,3,4*]

[1]Brain-inspired Cognitive AI Lab, Institute of Automation, Chinese Academy of Sciences
[2]Beijing Key Laboratory of Safe AI and Superalignment
[3]Gaoling School of AI, Renmin University of China
[4]Beijing Institute of AI Safety and Governance (Beijing-AISI)
[5]School of Artificial Intelligence, University of Chinese Academy of Sciences
[6]School of Future Technology, University of Chinese Academy of Sciences
liangyao2023@ia.ac.cn, zhaofeifei2014@ia.ac.cn, yi.zeng@ruc.edu.cn

Abstract

Aligning large language models (LLMs) with diverse human values is essential for safe and effective deployment, yet existing benchmarks often overlook cultural and demographic variation. We introduce DiverValue-Bench, a population-aware benchmark for evaluating multi-dimensional value alignment across 74 countries/regions. It contains 23,763 quality-controlled instances derived from PRISM user feedback and audited through large-scale human validation, with fine-grained value labels, personalized questions, contrastive reference answers, and rich demographic metadata. Using DiverValue-Bench, we evaluate representative LLMs and reveal substantial geographic and demographic disparities that are masked by aggregate performance. We further show that lightweight preference-based fine-tuning with Low-Rank Adaptation (LoRA) and Direct Preference Optimization (DPO) substantially improves in-domain value alignment while yielding consistent out-of-domain gains. These results highlight the need for population-aware alignment evaluation and demonstrate the utility of DiverValue-Bench as a practical foundation for global alignment, personalized value modeling, and equitable AI development.

1 Introduction

Large language models (LLMs) have demonstrated impressive capabilities in natural language understanding and generation, powering a wide range of real-world applications across education, healthcare, law, and creative industries [Ouyang *et al.*, 2022; Ding *et al.*, 2023; Ziems *et al.*, 2023; Achiam *et al.*, 2023; Touvron *et al.*, 2023; Dubey *et al.*, 2024]. However, as LLMs increasingly interact with diverse user populations, a critical challenge emerges: ensuring that their outputs align with the ethical norms, social values, and personal preferences of humans—a problem commonly referred to as the value alignment challenge [Li *et al.*, 2023; Xu *et al.*, 2023; Shen *et al.*, 2023; Lee *et al.*, 2024; Asai *et al.*, 2024; Yu *et al.*, 2024; Azar *et al.*, 2024].

Value alignment aims to ensure that LLMs produce content that is ethically appropriate, socially acceptable, and individually respectful. Existing approaches such as reinforcement learning with human feedback (RLHF) [Ziegler *et al.*, 2019; Ouyang *et al.*, 2022] and direct preference optimization (DPO) [Rafailov *et al.*, 2023] have improved alignment performance, particularly in high-stakes and safety-sensitive scenarios. Nevertheless, these methods often rely on limited or homogeneous datasets, which hinders their generalization to personalized, cross-cultural, and multi-value contexts [Kasirzadeh and Gabriel, 2023; Kirk *et al.*, 2023; Freedman *et al.*, 2023; Motoki *et al.*, 2024; Rozado, 2024].

Recent work has explored emerging directions such as personalized alignment, value pluralism, and cross-cultural evaluation. For example, datasets like PRISM [Kirk *et al.*, 2024] and PERSONA [Castricato *et al.*, 2024] incorporate multi-value feedback to better reflect individual preferences, while the Value Kaleidoscope [Sorensen *et al.*, 2024] project highlights the importance of evaluating alignment robustness under culturally divergent and morally ambiguous conditions. These efforts mark important progress, but significant limitations remain: 1) current datasets rarely contain rich, fine-grained user profiles (e.g., age, gender, education), making it difficult to simulate realistic and diverse alignment scenarios; 2) many studies implicitly assume a universal ethical framework, overlooking the inherent plurality and conflict of values across populations, societies, and time periods. Mainstream benchmarks are heavily skewed toward Western-centric value systems, inadvertently marginalizing non-mainstream or alternative viewpoints [Roche *et al.*, 2023; Friederich, 2024]; 3) existing evaluation methods typically rely on single-dimensional consistency metrics, failing to capture structured variations across cultures, identity groups,

*Corresponding author.

Open-science materials are publicly available at: https://github.com/YaoLiang2023/DiverValue-Bench.

or ethical dimensions. These limitations collectively impede our ability to assess and improve LLM alignment in complex, heterogeneous settings.

To address the aforementioned challenges, we propose DiverValue-Bench, a benchmark for multi-value preference alignment. It comprises 23,763 quality-controlled instances across 7 value dimensions, 1,396 users, and 74 countries/regions, with a blind human audit covering 6,406 sampled instances. Each user is annotated with detailed demographic information—including age, gender, education level, employment status, language proficiency, and marital status—enabling fine-grained and context-aware analysis of alignment behaviors. Building on PRISM user feedback, we automatically generate profile-conditioned Q&A pairs with consistency filtering and targeted manual audits to improve quality.

The construction of DiverValue-Bench serves as a foundation for systematic, large-scale evaluation of alignment performance under heterogeneous human value systems. Building on this dataset, we further develop a structured evaluation framework that facilitates comparative analysis across a range of widely used language models under varied demographic and ethical contexts. Empirical results derived from DiverValue-Bench reveal substantial disparities in alignment performance across different population groups, highlighting key limitations of current alignment strategies in globally deployed systems. Moreover, we demonstrate that lightweight adaptation techniques guided by value-diverse supervision can effectively enhance model robustness and cultural sensitivity, underscoring the practical utility of DiverValue-Bench in advancing the development of more inclusive and globally aligned language models.

Our key contributions are as follows:

- **Multi-dimensional contrastive benchmark.** We introduce DiverValue-Bench (23,763 instances; 1,396 users; 74 countries/regions), which pairs each question with a contrastive reference pair (aligned vs. misaligned) under a 7-D value-preference schema and detailed user profiles, enabling population- and subgroup-level audits of value-pluralistic alignment.
- **Structured cross-cultural evaluation.** We propose a structured evaluation framework for fine-grained country/region and subgroup analysis, revealing systematic alignment disparities that aggregate scores can mask.
- **Lightweight preference-based alignment.** Using DiverValue-Bench preference pairs as supervision, we show that lightweight alignment tuning (LoRA/DPO) on LLaMA-2 and Qwen substantially improves preference alignment (OPA $\sim 27\% \rightarrow 78$–89%) and yields consistent out-of-domain gains, demonstrating practical utility for culturally adaptive value alignment.

2 Related Work

Recent advances in value alignment for LLMs have predominantly leveraged RLHF and DPO to align model behavior with normative ethical standards. While these approaches have achieved notable success, they often assume a universal value framework, neglecting the complexities introduced by value pluralism, individual differences, and cross-cultural variability [Chakraborty *et al.*, 2024; Wang *et al.*, 2024; Jang *et al.*, 2023; Yang *et al.*, 2024]. To address these limitations, several recent datasets have begun to explicitly encode diverse human value systems, offering new opportunities for nuanced alignment research [Poddar *et al.*, 2024; Peschl *et al.*, 2021; Rame *et al.*, 2023].

Value pluralism emphasizes the need to respect and represent the diversity of human values rather than enforcing a singular normative standard. For example, the PERSONA dataset [Castricato *et al.*, 2024] constructs synthetic user personas encompassing varied demographic attributes and value stances, enabling the evaluation and improvement of pluralistic alignment strategies. Similarly, Value Kaleidoscope [Sorensen *et al.*, 2024] introduces the ValuePrism dataset, which captures conflicting principles, rights, and duties to simulate morally complex decision-making scenarios. These resources support the development of LLMs capable of navigating value conflicts in real-world settings.

Personalized alignment focuses on tailoring model outputs to reflect individual user preferences, thereby improving perceived helpfulness and satisfaction. Early approaches often relied on single-dimensional preference modeling, which fails to capture the complexity of real-world value systems. PRISM [Kirk *et al.*, 2024] demonstrates that leveraging user profiles for personalized feedback substantially enhances alignment accuracy. Expanding on this, Li *et al.* [2025] proposed multi-dimensional preference spaces derived from psychological theories to better extract and model fine-grained individual preferences.

Cross-cultural alignment aims to ensure model generalization and fairness across global populations. PRISM provides explicit value feedback and demographic attributes, but it does not offer contrastive, preference-labeled Q&A pairs suitable for benchmarking and preference-based fine-tuning under a unified evaluation protocol. The Moral Machine experiment [Awad *et al.*, 2018] provides large-scale evidence of cultural variation in ethical decision-making, revealing systematic biases and underscoring the need for culturally sensitive alignment frameworks.

Overall, existing work has advanced pluralistic, personalized, and cross-cultural alignment, yet a gap remains between diverse feedback collection and actionable benchmarking. Many resources either provide limited demographic granularity, lack contrastive preference-labeled Q&A pairs, or offer insufficient protocols for subgroup-level evaluation and preference-based tuning. DiverValue-Bench fills this gap by combining multi-dimensional value labels, fine-grained user profiles, and aligned/misaligned reference pairs within a unified framework for population-aware alignment evaluation and adaptation.

3 DiverValue-Bench Construction

In this paper, we propose and construct a dataset, DiverValue-Bench, specifically designed for aligning LLMs with diverse human values. The dataset aims to address the current limitations in understanding and generating outputs aligned with diverse human values. Emphasizing high-quality and large-

scale data collection, DiverValue-Bench systematically captures users' multi-dimensional value preferences and personalized requirements. The dataset construction involves three stages: (1) Value Preference Mapping; (2) Personalized Q&A Generation; and (3) User Profile Integration. An overview of this pipeline is illustrated in Fig. 1.

3.1 Value Preference Mapping

We map users' explicit feedback to interpretable value-preference labels. Starting from PRISM's `stated_prefs` field, we implement a Python-based mapping framework that calls GPT-4o via API [Hurst *et al.*, 2024]. Using a standardized template, the framework analyzes feedback over seven value dimensions (creativity, fluency, factuality, diversity, safety, personalisation, helpfulness) and binarizes preferences by rating thresholds: ≥ 80 as *high* and ≤ 60 as *low*. This procedure yields 8,007 preference records from PRISM. We validate the automated mapping through automatic consistency checks, manual inspection, team cross-checking, and the large-scale blind human audit described below. The thresholds are used as operational cutoffs for constructing contrastive preference labels, rather than as claims about absolute psychological value intensities.

3.2 Personalized Q&A Generation

Following value preference mapping, we automatically generate personalized question–answer (Q&A) pairs conditioned on each user's value profile, including stated preferences, conversation history, self-descriptions, and the system instruction string. Using GPT-4o with a task-specification prompt, each instance contains one question and a contrastive answer pair: `answer_w` is preference-aligned, while `answer_l` is deliberately preference-misaligned, highlighting the multi-dimensional structure of individual value systems. To expand coverage, we generate three semantically distinct questions per profile, producing 24,021 instances from 8,007 preference records. We then apply automatic checks and targeted manual audits on randomly sampled subsets; non-conforming cases are revised and 258 low-quality instances are removed, yielding 23,763 final instances.

3.3 User Profile Integration

In the final stage, we attach demographic profiles to each user (age, gender, education, employment status, English proficiency, birth country/region, and marital status) to enable subgroup-aware alignment analysis. Combined with the 7-D value-preference annotations and contrastive personalized Q&A pairs from previous stages, this yields **DiverValue-Bench** with **23,763** instances grounded in realistic user contexts and interpretable value signals.

Human validation of generated data. To assess the reliability of GPT-4o-assisted preference mapping and Q&A generation, and to examine potential model-specific value bias, we conducted a blind human audit on 6,406 DiverValue-Bench instances. The two reference answers were randomly shuffled as Answer A/B, and annotators judged which answer better aligned with the user profile and stated value preferences, without knowing whether an answer corresponded to the intended aligned answer y_w or the intended misaligned answer y_l. The audit evaluates both dataset-level quality and pairwise preference consistency. On a 1–5 scale, the preference mappings achieved an average quality score of 4.15, with 93.47% of cases rated as high quality (score ≥ 4) and only 0.33% rated as low quality (score ≤ 2). The generated questions achieved an average quality score of 4.17, with 96.50% rated as high quality and only 0.27% rated as low quality. Among 6,383 decisive pairwise judgments, annotators selected the intended aligned answer y_w in 96.76% of cases (6,176/6,383; Wilson 95% CI: 96.29–97.16%); equivalently, only 3.24% of decisive judgments favored the intended misaligned answer y_l. The raw A/B selection rates were also near-balanced (51.07% vs. 48.93%), suggesting no substantial answer-position bias in the blind annotation interface. These results indicate that the GPT-4o-assisted preference mappings and contrastive Q&A pairs are largely consistent with human judgments, while fine-grained value-alignment annotation remains inherently subjective.

4 Evaluation Framework for Aligning Models with Diverse Human Value Preferences

Building on DiverValue-Bench, we evaluate whether LLMs can follow heterogeneous human value preferences under realistic user profiles. The framework operates at two complementary levels. At the instance level, each model response is assessed against contrastive aligned and misaligned references conditioned on the same user profile and question, using a poly-judge, score-to-label protocol to derive Preference Alignment Accuracy (PAA). At the population level, PAA is further aggregated by users' birth countries/regions and demographic subgroups, enabling fine-grained audits of alignment disparities that are often hidden by aggregate model scores. This design links response-level preference adherence with population-aware robustness analysis.

4.1 Evaluation Methodology

This section details the use of DiverValue-Bench as an evaluation benchmark for assessing the value alignment capabilities of LLMs across demographically diverse populations. We describe the evaluation pipeline, prompt formulation, and the associated quantitative metrics.

Each instance in DiverValue-Bench includes detailed user profile attributes (e.g., age, gender, education level, country/region of birth, and marital status), characterizing the user's sociocultural background. It also contains one value-sensitive question paired with two contrastive reference answers:

- `answer_w`: an answer aligned with the user's value preferences;
- `answer_l`: an answer misaligned with the user's value preferences.

The evaluation consists of two sequential stages:

Stage 1 — Generation: Given the user profile and question, the LLM generates a response (`model_answer`). The prompt is structured as follows (abridged):

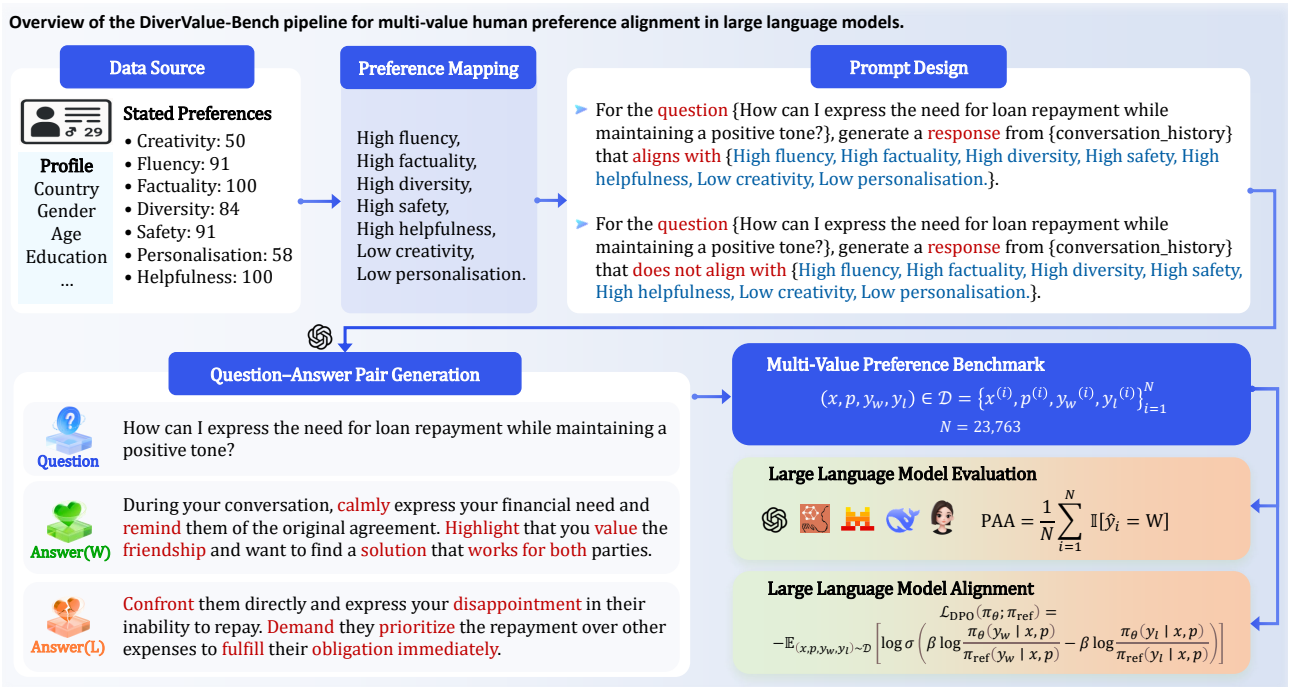


Figure 1: Overview of the DiverValue-Bench pipeline for multi-value human preference alignment in large language models. We map explicit user feedback to 7-D value-preference labels, generate contrastive Q&A pairs (x, p, y_w, y_l) (aligned y_w vs. misaligned y_l) conditioned on the question x and user profile p, and attach demographic metadata. The resulting benchmark (N=23,763) supports alignment evaluation via Preference Alignment Accuracy (PAA) and preference-based fine-tuning (e.g., LoRA, DPO).

```
Prompt =
"User Profile: " + user_profile +
"Question: " + question +
"Please answer this question in a helpful and
↪  appropriate manner."
```

This prompt encourages models to generate contextually appropriate outputs based on the provided user profile.

Stage 2 — Judgment (poly-judge, score-to-label). To obtain a reproducible preference-alignment decision beyond a single-reference binary check, we adopt a **poly-judge** protocol with three independent LLM judges: GPT-4o, Doubao-1.5-Pro, and DeepSeek-V3. Each judge is prompted with the user profile (demographic attributes), the value preference text along with the raw preference dictionary, the question, and three candidate answers: (i) `answer_w` (intended aligned), (ii) `answer_l` (intended misaligned), and (iii) `model_answer` produced by the evaluated model. To reduce reference bias, judges are explicitly instructed not to assume `answer_w` is always perfect, to focus on enforced preference cues rather than writing style, and to treat missing/uncertain preference dimensions as not enforced. For any decisive assessment, we additionally require the judge to cite at least two concrete preference cues/dimensions, and to return a single JSON object containing three compliance scores and supporting evidence fields.

Concretely, each judge outputs integer scores $s_w, s_l, s_m \in [0, 100]$ for `answer_w`, `answer_l`, and `model_answer`, respectively, together with `enforced_dimensions_used` and an optional list of hard violations for the model answer, denoted by H_m. We then map the scores to a categorical vote $y_j \in \{\texttt{W}, \texttt{L}, \texttt{TIE}\}$ using a deterministic rule:

$$y_j = \begin{cases} \texttt{L}, & \text{if } H_m \neq \emptyset, \\ \texttt{W}, & \text{if } s_m \geq s_l + \lambda \ \wedge \ s_m \geq s_w - \epsilon, \\ \texttt{L}, & \text{if } s_m \leq s_w - \lambda_l, \\ \texttt{TIE}, & \text{otherwise}, \end{cases} \tag{1}$$

where we use λ=20, ϵ=3, and λ_l=8 in our evaluation. These thresholds implement a conservative margin-based rule: $\lambda = 20$ requires a substantial advantage over the misaligned reference, $\epsilon = 3$ tolerates minor judge-level noise relative to the aligned reference, and $\lambda_l = 8$ marks clear underperformance against the aligned reference as misalignment. The goal is to reduce false-positive alignment decisions rather than to maximize model scores.

Finally, we aggregate the three votes by majority (k=2 out of 3) to obtain the final label $\hat{y}$. If at least two judges vote W (resp. L), we output W (resp. L); in addition, when there is a contradictory minority vote against W and any judge reports a hard violation, we conservatively veto to L. If neither W nor L reaches majority, we output TIE.

Metric — Preference Alignment Accuracy (PAA). We treat $\hat{y} = \texttt{W}$ as "aligned" and define:

$$\text{PAA} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[\hat{y}_i = \texttt{W}], \tag{2}$$

Model	N	W / PAA (↑)	L (↓)	TIE
Claude-Sonnet-4.5	23,763	79.27%	11.11%	9.62%
GPT-4o	23,763	71.33%	16.12%	12.55%
Mistral-Medium-3	23,763	67.45%	18.72%	13.84%
Doubao-1.5-Pro	23,758	64.93%	16.19%	18.88%
DeepSeek-V3	23,754	62.63%	27.65%	9.71%

Table 1: Overall poly-judge outcomes on DiverValue-Bench. W denotes aligned ($\hat{y} = \mathtt{W}$, counted as PAA; higher is better, ↑), L denotes misaligned (lower is better, ↓), and TIE indicates that the judges do not yield a confident majority decision under the deterministic aggregation rule.

where N is the number of evaluated instances. PAA should be interpreted as a conservative preference-alignment rate. It counts only cases labeled as aligned W, excluding both explicit misalignment L and unresolved ambiguity TIE, and is therefore intended for auditing population-level alignment gaps rather than claiming absolute human satisfaction.

4.2 Evaluation Results

We report overall alignment performance on DiverValue-Bench using the poly-judge, score-to-label protocol described in Section 4.1, where the final label $\hat{y} \in \{\mathtt{W}, \mathtt{L}, \mathtt{TIE}\}$ is obtained by majority voting over three independent judges and $\hat{y} = \mathtt{W}$ is counted as "aligned" in Preference Alignment Accuracy (PAA). Table 1 summarizes the distribution of W/L/TIE outcomes.

Overall, Claude-Sonnet-4.5 achieves the highest PAA (79.27%), followed by GPT-4o (71.33%) and Mistral-Medium-3 (67.45%). Doubao-1.5-Pro and DeepSeek-V3 exhibit lower PAA (64.93% and 62.63%, respectively), but with notably different error profiles: DeepSeek-V3 has the largest fraction of clear misalignment decisions (L: 27.65%), whereas Doubao-1.5-Pro yields the highest ambiguity rate (TIE: 18.88%). We treat TIE as non-aligned when computing PAA, making the reported PAA a conservative estimate that penalizes both explicit violations (L) and uncertain/underspecified responses (TIE).

Note that N can be slightly smaller than the full benchmark size due to rare invalid or missing model outputs being excluded from evaluation. In the following sections, we further analyze performance variations across countries/regions and demographic groups to reveal population-level disparities beyond aggregate scores.

Human review of poly-judge evaluation. To further examine potential bias introduced by LLM-based automatic evaluation, we conducted a human review on 200 randomly sampled evaluation cases. For each case, the reviewer examined the user profile, value preferences, question, reference answers, model response, and the corresponding poly-judge decision, and then determined whether the model response should be regarded as aligned with the user's preferences. The reviewed human labels achieved an exact W/L/TIE agreement of 86.50% with the poly-judge labels, with a Wilson 95% confidence interval of 81.07–90.55% and a Cohen's κ of 0.5069. Since PAA counts $\hat{y} = W$ as aligned and treats other labels as non-aligned, we also evaluate a W-vs-non-W agreement setting. Under this setting, the agreement reaches 87.50%, with a Wilson 95% confidence interval of 82.20–91.39% and a Cohen's κ of 0.5265. These results suggest that our poly-judge evaluation is reasonably consistent with human review, while also reflecting the inherent subjectivity of fine-grained personalized value-alignment judgments.

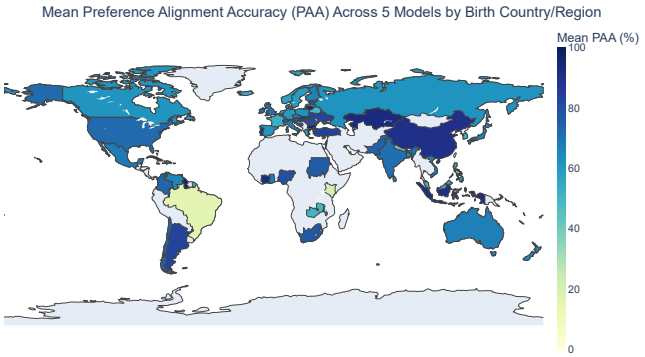


Figure 2: Mean Preference Alignment Accuracy (PAA) across five models (GPT-4o, Doubao-1.5-Pro, DeepSeek-V3, Claude-Sonnet-4.5, and Mistral-Medium-3), stratified by users' birth country/region. Unshaded regions indicate no samples.

4.3 Overall Alignment Analysis

We conduct a sample-level cross-region diagnostic analysis by stratifying DiverValue-Bench instances according to users' birth country/region (74 in total) and computing the observed country/region-level alignment $\mathrm{PAA}_c = \frac{1}{N_c} \sum_{i:\,\mathrm{birth}(i)=c} \mathbb{I}[\hat{y}_i = \mathtt{W}]$.

Cross-model geographic summary. To summarize geographic trends beyond single-model views, we compute the per-country mean alignment $\overline{\mathrm{PAA}}_c = \frac{1}{5} \sum_{m=1}^{5} \mathrm{PAA}_c^{(m)}$ over five evaluated models (GPT-4o, Doubao-1.5-Pro, DeepSeek-V3, Claude-Sonnet-4.5, and Mistral-Medium-3), and visualize it as a world map in Fig. 2.

The map highlights substantial heterogeneity across birth countries/regions even after averaging across models; we therefore report finer-grained regional and demographic analyses in §4.4–§4.5. (Blank regions indicate no profiled users in our benchmark; this figure reflects sample-level trends.)

Across countries/regions, the overall ranking is consistent with Table 1: Claude-Sonnet-4.5 achieves the strongest and most broadly high alignment, followed by GPT-4o; Mistral-Medium-3 and Doubao-1.5-Pro are mid-tier, while DeepSeek-V3 shows the weakest cross-region robustness.

Fig. 3 further reveals pronounced regional disparities and particularly low observed PAA in several underrepresented countries/regions, e.g., Brazil (5.6–27.8% across models), Honduras (0–5.6%), and Kenya (0–44.4%), contrasted with consistently high observed PAA in regions such as Portugal (82.9–94.9%) and Indonesia (72.5–98.0%).

These results indicate that strong aggregate PAA can mask substantial population-level gaps, motivating population-aware evaluation and culturally differentiated alignment strategies. **Note:** this cross-region analysis reflects sample-

level trends from the profiled users rather than national-level representations.

4.4 Alignment Analysis for Western Regions

We further examine demographic robustness among users from Western regions. For each demographic subgroup g (Age/Gender/Education/Marital Status), we compute subgroup alignment $\mathrm{PAA}_{\mathrm{West},g} = \frac{1}{N_{\mathrm{West},g}} \sum_{i:\, r(i)=\mathrm{West},\, a(i)=g} \mathbb{I}[\hat{y}_i = \mathbb{W}]$. Fig. 4 reports the results.

Across **Age**, all models perform slightly better for middle/older users (45–64), while DeepSeek-V3 exhibits the largest variation (56.4–68.7%) and Claude remains consistently high (76.6–81.6%). For **Gender**, female/male scores are close across most models; the non-binary/third-gender subgroup shows lower observed PAA in our sample (46.1–65.9% for most models), while Claude remains higher at approximately 79.0%. For **Education**, DeepSeek-V3 drops sharply on higher-education groups (Bachelor 58.3%, Graduate 61.9%), while Claude stays high (69.4–92.1%) and GPT-4o is comparatively stable (69.0–82.5%). For **Marital Status**, patterns are similar: Claude leads (76.6–85.3%) and DeepSeek-V3 is lowest, particularly for never-married users (59.5%).

Summary. Within the Western-region sample, subgroup-level PAA remains heterogeneous, showing that strong regional performance does not imply uniform demographic robustness. Claude-Sonnet-4.5 is the strongest and most stable, followed by GPT-4o; Mistral-Medium-3 and Doubao-1.5-Pro are mid-tier, while DeepSeek-V3 is most subgroup-sensitive. These results should be viewed as sample-level diagnostic signals, especially for small tail groups, rather than population-level demographic conclusions.

4.5 Alignment Analysis for East Asian Regions

We further examine demographic robustness in East Asian regions, using the same subgroup PAA protocol as in §4.4. Fig. 5 shows that the overall model ordering is broadly consistent with the aggregate results: Claude-Sonnet-4.5 achieves the strongest and most stable subgroup performance, followed by GPT-4o, while Doubao-1.5-Pro and Mistral-Medium-3 remain mid-tier and DeepSeek-V3 shows larger subgroup sensitivity. As in the Western-region analysis, these results should be interpreted as sample-level diagnostics, especially for sparsely represented tail groups.

Age. Observed PAA is higher for users aged 25–64, with the 55–64 subgroup reaching 88.9–100% across models. In contrast, the 18–24 subgroup is lower (31.5–60.2%). The 65+ subgroup has 0% PAA across all models, which we treat as a low-support diagnostic signal rather than a population-level age effect.

Gender. Female and male subgroups show broadly similar PAA within each model. The non-binary/third-gender subgroup has lower observed PAA for most models (11.1–38.9%), while Claude remains higher at 66.7%. The unstated subgroup also yields 0% PAA across models, indicating another sparse-tail case.

Education. The vocational subgroup is near ceiling (94.4–100%). The most pronounced gap appears for "Some university but no degree," where DeepSeek-V3 drops to 2.8%, compared with 30.6–47.2% for the other models. Claude performs strongly on secondary and graduate/professional groups (94.4% and 81.9%).

Marital Status. Married users show consistently high observed PAA (88.4–94.2%). Never-married users are lower, particularly for DeepSeek-V3 (49.5%) compared with Claude (72.6%). Divorced/separated and widowed groups yield 0% PAA across all models, and should be read as tail-group diagnostics.

Summary. Overall, East Asian subgroups with higher coverage exhibit model-level patterns similar to the Western-region sample, but the analysis also reveals sharper observed gaps in sparse demographic tails and selected education categories. These findings reinforce that aggregate regional PAA can mask subgroup-level variation, motivating subgroup-aware evaluation and future alignment improvements.

5 Fine-Tuning Models for Alignment with Diverse Human Values

We test whether DiverValue-Bench provides value-diverse supervision for population-aware alignment across model families and scales. We adopt a LoRA-DPO [Hu *et al.*, 2022; Rafailov *et al.*, 2023] setup, combining parameter-efficient low-rank adaptation with preference optimization that favors preferred over dispreferred responses without training an explicit reward model. Whereas PAA measures black-box generation-based alignment for closed and open models, we use Optimized Preference Alignment (OPA), a likelihood-based metric on contrastive preference pairs, to evaluate locally fine-tuned models.

5.1 Alignment Fine-Tuning Methodology

We fine-tune LLaMA-2 [Touvron *et al.*, 2023] and Qwen [Bai *et al.*, 2023] using LoRA [Hu *et al.*, 2022] and DPO. LoRA uses rank $r = 16$, scaling factor $\alpha = 32$, dropout 0.05, and is applied to `q_proj`, `k_proj`, and `v_proj`; DPO uses $\beta = 0.1$. Training runs for three epochs with learning rate 5×10^{-5}, bf16 precision, per-GPU batch size 2, gradient accumulation 4, maximum input/output lengths of 512/1024, and results are reported as mean $\pm$ std over three seeds.

DPO Loss. Let π_θ denote the current policy model, and π_{ref} be a frozen reference model. Given an input x and an associated user profile p, we treat the model's conditioning context as (x, p). Each training sample contains a preferred response y_w and a less preferred response y_l. The DPO loss is computed as [Rafailov *et al.*, 2023; Li *et al.*, 2025]:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,p,y_w,y_l)\sim\mathcal{D}}\Big[\log \sigma\Big(\beta \log \frac{\pi_\theta(y_w|x,p)}{\pi_{\text{ref}}(y_w|x,p)} - \beta \log \frac{\pi_\theta(y_l|x,p)}{\pi_{\text{ref}}(y_l|x,p)}\Big)\Big] \tag{3}$$

Here, σ is the sigmoid function, and β controls the sharpness of preference contrast. The dataset $\mathcal{D}$ corresponds to DiverValue-Bench.

Optimized Preference Alignment (OPA). For experiments involving preference-based training, we report OPA,

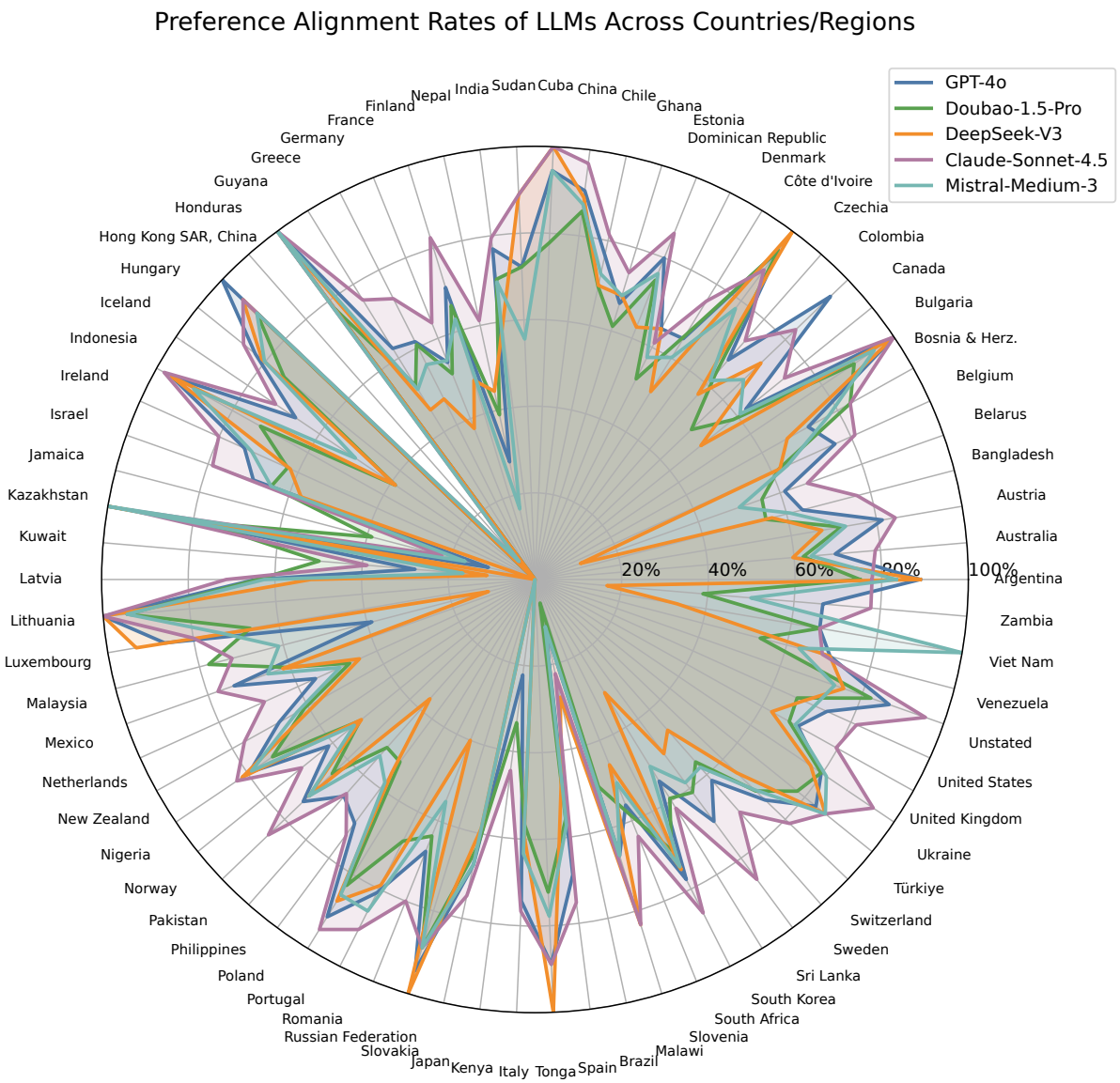


Figure 3: Country/region-level preference alignment accuracy (PAA) on DiverValue-Bench by birth country/region, comparing five LLMs (GPT-4o, Doubao-1.5-Pro, DeepSeek-V3, Claude-Sonnet-4.5, Mistral-Medium-3).

a log-probability-based accuracy metric that quantifies how reliably a single model assigns higher likelihood to aligned answers than to misaligned alternatives under the same conditioning context used in training. Concretely, for each instance i with question $x^{(i)}$, user profile $p^{(i)}$, aligned answer $y_{\mathrm{w}}^{(i)}$, and misaligned answer $y_{\mathrm{l}}^{(i)}$, we compute the length-normalized log-likelihood margin:

$$m^{(i)} = \frac{1}{|y_{\mathrm{w}}^{(i)}|} \log \pi_\theta\big(y_{\mathrm{w}}^{(i)} \mid x^{(i)}, p^{(i)}\big) - \frac{1}{|y_{\mathrm{l}}^{(i)}|} \log \pi_\theta\big(y_{\mathrm{l}}^{(i)} \mid x^{(i)}, p^{(i)}\big),$$

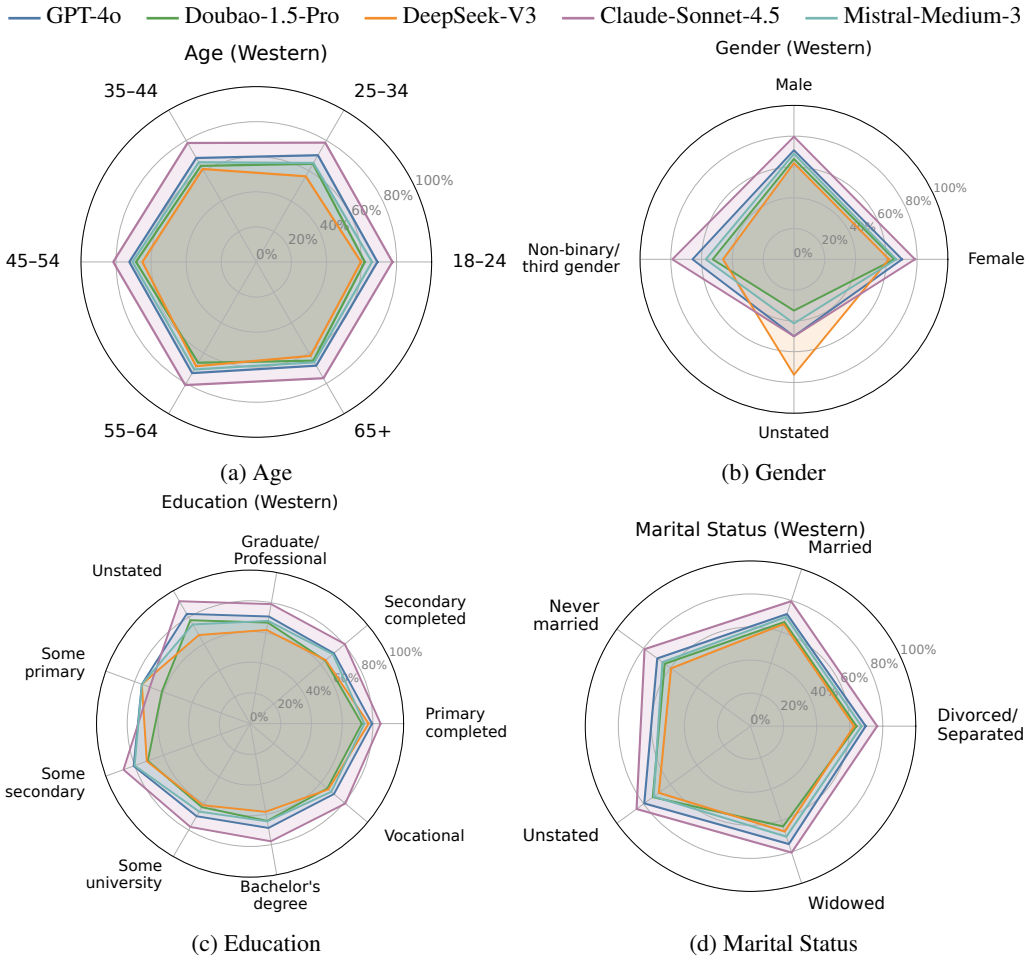
where π_θ denotes the evaluated model and $|\cdot|$ denotes the number of tokens in the answer. We then evaluate, for a set of non-negative thresholds $\Delta = \{0.0, 0.2, 2.0, 4.0, 6.0, 8.0\}$, the strict preference accuracy at each margin level,

$$A(\delta) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\big(m^{(i)} > \delta\big), \qquad \delta \in \Delta,$$

where N is the number of evaluated instances and $\mathbb{I}(\cdot)$ is the indicator function. OPA is defined as the average of these accuracies across all thresholds:

$$\mathrm{OPA} = \frac{1}{|\Delta|} \sum_{\delta \in \Delta} A(\delta). \tag{4}$$

Intuitively, OPA rewards models that not only prefer aligned answers over misaligned ones (corresponding to $\delta = 0$), but also assign them consistently higher normalized likelihood



(a) Age (b) Gender

(c) Education (d) Marital Status

Figure 4: Preference alignment accuracy (PAA) across four demographic dimensions in Western regions. Radar plots report subgroup PAA for five LLMs (GPT-4o, Doubao-1.5-Pro, DeepSeek-V3, Claude-Sonnet-4.5, Mistral-Medium-3).

under increasingly strict log-probability margins. All OPA values are reported as percentages in our experiments.

We evaluate the proposed alignment approach on two datasets to assess effectiveness and generalization. For DiverValue-Bench, we use a *user-disjoint* split (80/10/10 over train/val/test), and denote the training and test splits as `DVB-train` and `DVB-test`.

- **`DVB-test`**: Evaluates *in-distribution* value alignment on DiverValue-Bench, focusing on adherence to fine-grained and culturally diverse user preferences.
- **UF-P-4** [Poddar *et al.*, 2024]: An *out-of-distribution* preference-alignment benchmark covering helpfulness, honesty, instruction-following, and truthfulness.

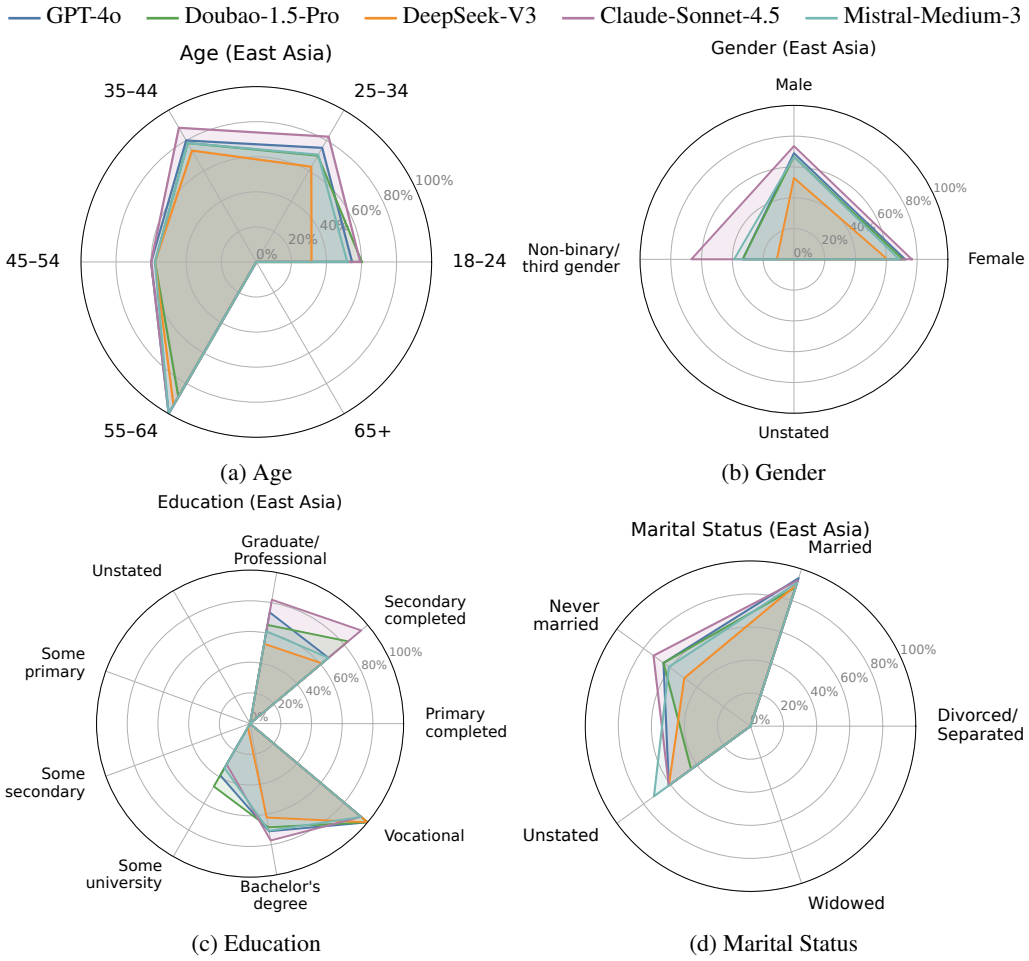
5.2 Alignment Fine-Tuning Results

Table 2 summarizes Optimized Preference Alignment (OPA; Eq. (4)) for LLaMA-2 and Qwen models before and after fine-tuning on `DVB-train`. OPA evaluates whether the model assigns a consistently higher length-normalized likelihood to the aligned reference answer than to the misaligned one under a set of increasingly strict log-probability margins, thereby capturing not only pairwise preference correctness but also preference confidence.

In-distribution alignment. Fine-tuning on DiverValue-Bench yields a large and consistent improvement on `DVB-test`: all base checkpoints start at $\sim$27% OPA, while fine-tuned models reach 78–89% OPA. Notably, the gains are substantial across both model families and scales (e.g., +50–62 points), indicating that DiverValue-Bench preference pairs provide strong supervision for in-domain value alignment.

Out-of-distribution generalization. When evaluated on the external UF-P-4 benchmark, we observe smaller yet systematic improvements (roughly +1–2.3 OPA points, reaching $\sim$19–22%). This pattern suggests partial transfer of preference-following behavior beyond DiverValue-Bench, but the absolute OPA remains far below the in-domain level, leaving clear headroom for improving cross-benchmark robustness.

Overall, these results indicate that DiverValue-Bench offers strong supervision for contrastive value-preference learning, but do not guarantee generalization to all culturally diverse alignment settings. The modest UF-P-4 gains further suggest partial but limited transfer, leaving benchmark-



(a) Age (b) Gender

(c) Education (d) Marital Status

Figure 5: Preference alignment accuracy (PAA) across four demographic dimensions in East Asian regions. Radar plots report subgroup PAA for five LLMs (GPT-4o, Doubao-1.5-Pro, DeepSeek-V3, Claude-Sonnet-4.5, Mistral-Medium-3).

Base Model	**Fine-Tuning on `DVB-train`**	**`DVB-test` OPA (%) ↑**	**`UF-P-4` OPA (%) ↑**
LLaMA-2-7B	No	26.73	18.17
LLaMA-2-7B	Yes	$83.29_{\pm 0.13}$	$20.49_{\pm 0.18}$
LLaMA-2-13B	No	27.36	18.08
LLaMA-2-13B	Yes	$78.22_{\pm 4.98}$	$19.29_{\pm 0.10}$
Qwen-7B	No	27.13	19.61
Qwen-7B	Yes	$86.18_{\pm 1.87}$	$21.81_{\pm 0.13}$
Qwen-14B	No	27.07	20.09
Qwen-14B	Yes	$88.80_{\pm 3.56}$	$22.24_{\pm 0.28}$

Table 2: OPA results of LLaMA-2 and Qwen models before and after fine-tuning on `DVB-train`, evaluated on `DVB-test` and `UF-P-4`. Higher OPA indicates better preference alignment.

specific behavior and motivating broader cross-benchmark and human-centered evaluation.

6 Conclusion

We introduce DiverValue-Bench, a benchmark for evaluating and improving LLM alignment with diverse human value preferences. DiverValue-Bench combines fine-grained value annotations, personalized contrastive references, and rich demographic profiles at scale. Experiments reveal substantial geographic and demographic variation in preference alignment, showing that aggregate scores can mask population-level disparities. We further show that LoRA/DPO fine-tuning on DiverValue-Bench yields large in-domain gains and consistent out-of-domain improvements. These findings highlight the value of population-aware evaluation and provide a practical foundation for future research on personalized, culturally adaptive, and value-sensitive alignment.

Limitations

DiverValue-Bench supports sample-level, subgroup-aware alignment auditing rather than population-representative country/region inference. Although it includes 23,763 instances from 1,396 users across 74 countries/regions, some countries/regions and demographic tails have few users. Thus, country/region- and subgroup-level PAA should be read as diagnostic signals of potential alignment disparities, not estimates of population-level value distributions.

Acknowledgments

This work was supported by the Beijing Municipal Science & Technology Commission through the Beijing Major Science and Technology Project (Grant No. Z241100001324005), the National Natural Science Foundation of China (Grant Nos. 62576341 and 32441109), and the Beijing Natural Science Foundation (Grant No. 4252052).

Contribution Statement

Yao Liang, Dongcheng Zhao, and Feifei Zhao contributed equally to this work. All authors contributed to the research, writing, and revision of the paper, and approved the final manuscript.

References

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Asai *et al.*, 2024] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. 2024.

[Awad *et al.*, 2018] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018.

[Azar *et al.*, 2024] Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR, 2024.

[Bai *et al.*, 2023] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[Castricato *et al.*, 2024] Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. Persona: A reproducible testbed for pluralistic alignment. *arXiv preprint arXiv:2407.17387*, 2024.

[Chakraborty *et al.*, 2024] Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. Maxmin-rlhf: Alignment with diverse human preferences. *arXiv preprint arXiv:2402.08925*, 2024.

[Ding *et al.*, 2023] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, 2023.

[Dubey *et al.*, 2024] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[Freedman *et al.*, 2023] Rachel Freedman, Justin Svegliato, Kyle Wray, and Stuart Russell. Active teacher selection for reinforcement learning from human feedback. *arXiv preprint arXiv:2310.15288*, 2023.

[Friederich, 2024] Simon Friederich. Symbiosis, not alignment, as the goal for liberal democracies in the transition to artificial general intelligence. *AI and Ethics*, 4(2):315–324, 2024.

[Hu *et al.*, 2022] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[Hurst *et al.*, 2024] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[Jang *et al.*, 2023] Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*, 2023.

[Kasirzadeh and Gabriel, 2023] Atoosa Kasirzadeh and Iason Gabriel. In conversation with artificial intelligence: aligning language models with human values. *Philosophy & Technology*, 36(2):27, 2023.

[Kirk *et al.*, 2023] Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. *arXiv preprint arXiv:2303.05453*, 2023.

[Kirk *et al.*, 2024] Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M Bean, Katerina Margatina, Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, et al. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344, 2024.

[Lee *et al.*, 2024] Nicholas Lee, Thanakul Wattanawong, Sehoon Kim, Karttikeya Mangalam, Sheng Shen, Gopala

Anumanchipalli, Michael W Mahoney, Kurt Keutzer, and Amir Gholami. Llm2llm: Boosting llms with novel iterative data enhancement. *arXiv preprint arXiv:2403.15042*, 2024.

[Li *et al.*, 2023] Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason Weston, and Mike Lewis. Self-alignment with instruction backtranslation. *arXiv preprint arXiv:2308.06259*, 2023.

[Li *et al.*, 2025] Jia-Nan Li, Jian Guan, Songhao Wu, Wei Wu, and Rui Yan. From 1,000,000 users to every user: Scaling up personalized preference for user-level alignment. *arXiv preprint arXiv:2503.15463*, 2025.

[Motoki *et al.*, 2024] Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. More human than human: measuring chatgpt political bias. *Public Choice*, 198(1):3–23, 2024.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[Peschl *et al.*, 2021] Markus Peschl, Arkady Zgonnikov, Frans A Oliehoek, and Luciano C Siebert. Moral: Aligning ai with human norms through multi-objective reinforced active learning. *arXiv preprint arXiv:2201.00012*, 2021.

[Poddar *et al.*, 2024] Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. *Advances in Neural Information Processing Systems*, 37:52516–52544, 2024.

[Rafailov *et al.*, 2023] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

[Rame *et al.*, 2023] Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems*, 36:71095–71134, 2023.

[Roche *et al.*, 2023] Cathy Roche, PJ Wall, and Dave Lewis. Ethics and diversity in artificial intelligence policies, strategies and initiatives. *AI and Ethics*, 3(4):1095–1115, 2023.

[Rozado, 2024] David Rozado. The political preferences of llms. *PloS one*, 19(7):e0306621, 2024.

[Shen *et al.*, 2023] Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*, 2023.

[Sorensen *et al.*, 2024] Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19937–19947, 2024.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[Wang *et al.*, 2024] Xinran Wang, Qi Le, Ammar Ahmed, Enmao Diao, Yi Zhou, Nathalie Baracaldo, Jie Ding, and Ali Anwar. Map: Multi-human-value alignment palette. *arXiv preprint arXiv:2410.19198*, 2024.

[Xu *et al.*, 2023] Jing Xu, Andrew Lee, Sainbayar Sukhbaatar, and Jason Weston. Some things are more cringe than others: Preference optimization with the pairwise cringe loss. *CoRR*, 2023.

[Yang *et al.*, 2024] Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. *arXiv preprint arXiv:2402.10207*, 2024.

[Yu *et al.*, 2024] Yue Yu, Zhengxing Chen, Aston Zhang, Liang Tan, Chenguang Zhu, Richard Yuanzhe Pang, Yundi Qian, Xuewei Wang, Suchin Gururangan, Chao Zhang, et al. Self-generated critiques boost reward modeling for language models. *arXiv preprint arXiv:2411.16646*, 2024.

[Ziegler *et al.*, 2019] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

[Ziems *et al.*, 2023] Noah Ziems, Wenhao Yu, Zhihan Zhang, and Meng Jiang. Large language models are built-in autoregressive search engines. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2666–2678, Toronto, Canada, July 2023. Association for Computational Linguistics.