# Steerable Cultural Preference Optimization of Reward Models

Anonymous TACL submission


## Abstract

It is essential for large language model (LLM) technology to serve many different cultural sub-communities in a manner that is acceptable to each community. However, research on LLM alignment has so far predominantly focused on predicting a unified response preference of annotators from certain regions. This paper aims to advance the development of alignment models with a more global outlook, that are able to accurately represent the preferences of sub-communities and do not exhibit excessive bias towards any of them. We focus on the development of reward models for this purpose and present a novel reward model training algorithm (SCPO) that can incorporate diverse cultural preferences in a balanced manner. Our method results in performance increases of the minority reward model of up to 7 points over the baseline model across two datasets, PRISM and GlobalOpinionQA, and across 7 countries. SCPO is up to 170% more training data-efficient than full-data finetuning of reward models. In addition, we perform analysis of bias by separately evaluating on the preference of subcommunities and show that excessive bias is mitigated via our weighting method.


## 1 Introduction

Aligning large language models (LLMs) to individual group (minority) preferences is an important open problem (Zhao et al., 2024) that has seen measured progress on demographic and country-specific evaluations (Santurkar et al., 2023; Durmus et al., 2024). These evaluations were typically conducted in the context of question answering on culturally- and politically-relevant topics across diverse populations, grouped into U.S. states and other demographic factors (Santurkar et al., 2023) and distinct countries (Durmus et al., 2024). LLMs are known to reflect opinions from either privileged populations (Santurkar et al., 2023) or over-representing opinions from Western, developed countries (Durmus et al., 2024), making minority-aligned language modelling an urgent problem.

Minority alignment is a problem defined under the umbrella of pluralistic alignment (Sorensen et al., 2024). Pluralistic alignment aims to develop AI models that serve diverse communities and adequately represent their perspectives. Sorensen et al. (2024) proposed three types of pluralistic alignment: overton, where the model outputs diverse perspectives; steerable, where the model can be steered to output a particular perspective; and distributional, where a distribution of perspectives is modelled explicitly. Our approach to minority alignment aims to build steerable reward models that are specific to a country's point of view.

Several recent alignment frameworks aim to model group preferences. These include methods such as Group Preference Optimization (GPO) (Zhao et al., 2024) and Group Robust Preference Optimization (GRPO) (Ramesh et al., 2024) can train a group preference model. GPO utilizes a separate fine-tuned transformer module on top of LLM to predict a group's preferences. This makes it not straightforward to integrate into general-purpose LLM alignment frameworks, such as reinforcement learning with human feedback (RLHF) (Schulman et al., 2017) or direct preference optimization (DPO) (Rafailov et al., 2024), as it has not been developed with this in mind. GRPO, on the other hand, works with a specific definition of "robustness" and minimizes the worst-case group loss. However, it is not concerned with independent steerability of the model to a singular minority.

Central to our approach is the use of a 'global'

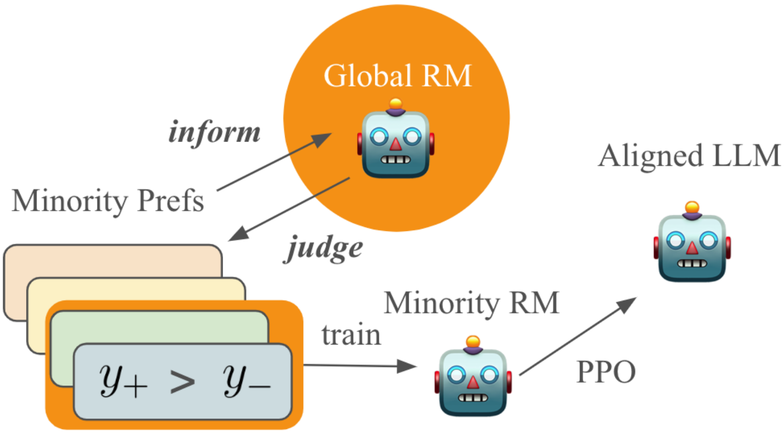


Figure 1: Overview of Steerable Cultural Preference Optimization (SCPO). The pipeline consists of: (1) a global reward model that scores minority preference pairs; (2) **Filtering** (Section 4.1): removing pairs where minority and global preferences agree; and (3) **Weighting** (Section 4.2): assigning lower training weights to extreme preferences. Yellow highlighting indicates our novel contributions. The right side shows downstream use in PPO-based RLHF. Note that GlobalRM may be same as starting checkpoint of RM for minority training.

reward model: an RM trained on broad preference data (e.g., OpenAssistant or Tülu 3). We leverage this global RM for two purposes:

1. **Identifying Cultural Distinctiveness**: Preference pairs where minority annotations disagree with global RM predictions represent genuinely distinctive cultural preferences. By filtering to retain only these disagreeing pairs, we focus training on what makes each culture unique rather than on universal preferences already captured by the global model.

2. **Measuring Extremeness**: Large disagreements (where the global RM assigns very different scores than minority annotations) may indicate extreme preferences that risk making the minority model overly biased. Our weighting scheme down-weights such extreme cases.

The global RM thus serves as a reference point representing ’mainstream’ or ’consensus’ preferences, enabling us to systematically identify and calibrate cultural deviations. Importantly, we do not assume the global RM is ’correct’. Rather, we use it as a tool to differentiate minority preferences from majority ones.

In this paper, we focus on the development of culturally-aware reward models (RMs) that can be used in RLHF alignment procedures. Specifically, we propose a novel method that utilizes a global reward model to identify culture-specific preference samples and present a weighted reward model training loss to conduct a multi-faceted balanced training of RMs. Our research questions are as follows:

1. **How do we ensure that minority reward models have balanced opinions?** While we want to reflect minorities’ opinions on LLM outputs, we want to simultaneously deemphasize undesired responses within a minority preference dataset. We design a two-tiered multi-faceted evaluation approach that utilizes distinct test sets to ensure we create a reward model with balanced opinions.

2. **Can we utilize global reward model preference scores for minority reward model training?** We devise a novel alignment method that utilizes open-source reward models that are not minority aligned. We utilize the scores given by these global reward models for both training and evaluation of the minority reward models.

3. **Which subsection of preference data is important for effective minority reward model training?** Some training preference pairs in minority preference data will agree with the global model, while other pairs will be different. We utilize the scores of the global reward models on certain preference pairs to either truncate or emphasize sections of pairwise preference data, and report observed performance tradeoffs.

We fine-tune two reward models (OpenAssistant and Tulu) on country-specific data from the PRISM dataset using our method and find filtering and weighting of the data, utilizing global reward model scores, is beneficial to the performance of our models on overall test set, while avoiding aligning to skewed preference.

## 2 Related Works

### 2.1 Prompt-Based Minority Alignment

Several works on cultural alignment utilize carefully crafted prompts to improve cultural responses in language models. Culture-Gen (Li et al., 2024b) uses open-source datasets and variations of system prompts (for instance, "*My neighbor is [nationality]. My neighbor is probably wearing...*,") to reveal the linguistic markers that influence generation. They determine the best system prompt to use across state-of-the-art models. Similarly, AlKhamissi et al. (2024) introduce Anthropological Prompting to ensure models reason critically on culturally sensitive topics. CultureLLM (Li et al., 2024a) performs cultural data augmentation using prompting techniques and fine-tuning LLMs on the generated data. However, these prompts are arbitrarily crafted with no rigorous testing to ensure they are optimal. Furthermore, by not using real cultural preference data, these approaches risk perpetuating preexisting biases in models' training data. All 3 approaches do not examine the extremity of outputs, which is important because they rely on models' skewed perceptions of minority culture.

### 2.2 Filtering Samples with Reward Models

Approaches such as reward ranked fine-tuning (RAFT) (Dong et al., 2023) and Supervised Iterative Learning from Human Feedback (SuperHF) (Mukobi et al., 2023), demonstrate the potential of using only the most valuable training examples to improve model performance. RAFT utilizes reward-based reranking by iteratively scoring samples via a reward function, filtering for high-reward examples, and fine-tuning the model using this subset. Similarly, SuperHF filters model-generated training data with a reward model and only uses high-reward synthetic data for fine-tuning. Both approaches demonstrate significant improvements by using a reward model to identify high-quality data. However, neither method targets minority alignment, accounts for preference pairs, or goes beyond basic reward thresholds for filtering.

### 2.3 Weighting Samples with Reward Models

Methods in weighting-based alignment, such as Online Preference Tuning (OPTune) (Chen et al., 2024b) and Mallows-DPO (Chen et al., 2024a), highlight the benefits of using reward models to prioritize certain samples. OPTune improves alignment by introducing a weighted DPO objective that emphasizes pairs with larger reward gaps, ensuring the model learns more from high-priority examples. Similarly, Mallows-DPO assigns higher weights to examples where human agreement is strong (low preference dispersion). Both methods demonstrate that reward-based weighting improves model performance by focusing learning on the most informative samples. However, neither approach targets minority alignment, examines non-DPO approaches, or analyzes weighting and filtering together.

## 3 Datasets

**PRISM** We primarily utilize PRISM (Kirk et al., 2024), a human feedback dataset for preference and value alignment of LLMs. PRISM is an LLM preference dataset comprising of controversal conversations between LLM and user across different countries. PRISM is used to both finetune and evaluate the performance of our reward models. We randomly split PRISM users into train and test sets using 8.5:1.5 user ratio, to ensure multi-turn data from conversations are not divided across the data splits. Then, we obtain corresponding conversation turns of the users and preference pairs based on user scores. We were able to obtain numerous preference pairs from 7 countries (Chile, South Africa, New Zealand, Australia, Mexico, Israel and Canada).

To fit our use cases, we re-structure both the `survey` data (which contains demographic information of the participants, as seen in Appendix B Table 9) and the `utterance` data (the content of the actual conversations between participants and LLMs and participant ratings, as seen in Appendix B Table 10) from PRISM.

**GlobalOpinionQA** Additionally, we use Anthropic's GlobalOpinionQA (Durmus et al., 2024) dataset to evaluate our country-specific reward models. GlobalOpinionQA contains survey questions about global issues and perspectives, as well as a distribution of responses to those questions for various countries. By providing the question as the prompt and each of the answer options as responses to the country-specific reward model, we can see if the relative ranking of the rewards given to each answer corresponds with the probability distribution of answers chosen by that country in GlobalOpinionQA.

## 4 Methodology

We develop a novel method (Fig. 1) of working with minority preferences in conjunction with global preferences, which consist of filtering and weighting stages. Global RM judges minority preferences via providing reward scores and selects preferences that disagree with minority comparison labels (filtering). Using the Global RM reward scores, each preference is weighted differently in weighted training loss, to ensure subtle differences are emphasized (weighting). Global RM can be reused from starting Tülu 3 and OpenAssistant models, while minority RM is a result of training the said models to given minority's preferences. See Fig. 2 for an example.

### 4.1 Filtering

We remove minority pairwise preferences from the training set if they agree with the global model preferences. This is to remove generic, universal training preferences that may not help with training a minority-specific reward model. Conversely, we retain preference pairs that disagree with the global model. By keeping only the minority pairwise preferences that disagree with the global model preferences, we aim to streamline the training of minority reward models by utilizing only necessary data to achieve greater data efficiency. In practice, about one half of training data is left after filtering, achieving 170% data efficiency. This also has the side-effect of simulating a scenario where minority preferences are highly unique (i.e. 90% of the preferences disagree with the global consensus).

We utilize the Bradley-Terry model (Bradley and Terry, 1952) for our filtering algorithm. We retain preference pairs where the Global RM disagrees with the minority label (indicating a unique cultural preference) and discard pairs where the Global RM already aligns with the minority label (indicating a generic preference). Our filtering algorithm is as follows:

$$p_{\text{glo}}(y_+ \succ y_-|x) = \frac{e^{r_{\text{glo}}(x,y_+)}}{e^{r_{\text{glo}}(x,y_+)} + e^{r_{\text{glo}}(x,y_-)}} < \tau \quad (1)$$

Per minority preference annotations, $y_+$ is the preferred response in the pair and $y_-$ is the dispreferred response in the pair. $p_{\text{glo}}(y_+ \succ y_-|x)$ is the probability that corresponds to the global model preferring $y_+$ data instead of $y_-$. $r_{\text{glo}}$ is global reward model that produces a score. $\tau$ is a $0 \leq \tau \leq 1$ threshold for subset selection of preference data.

### 4.2 Weighted RM training loss

We define the 'extremeness' of a preference pair $(y_+, y_)$ as the degree of disagreement between the minority annotation and the global RM's preference. Formally, a preference is extreme when $p_{\text{glo}}(y_+ \succ y_-|x)$ is high, meaning the global model strongly prefers the response that minority annotators rejected. We distinguish extremeness from related concepts:

- **Harmful content** - Responses containing toxic language, hate speech, or unsafe recommendations. While extreme preferences may correlate with harmful content (see Table 2), extremeness is defined purely by disagreement magnitude, not content analysis.
- **Cultural distinctiveness** - Preferences can be culturally distinctive without being extreme—e.g., preferring formal vs. informal language styles.
- **Annotator error** - Some extreme preferences may reflect inconsistent annotations rather than genuine cultural differences.

Our weighting scheme does not classify content as 'good' or 'bad' but rather modulates training influence based on disagreement magnitude, allowing the model to learn from distinctive preferences while reducing influence of outliers.

We develop a novel training loss that inversely assigns weights to the preferences pairs according to their extremeness. Thus, less extreme preference data is weighted more highly than more extreme preference data. (See Table 1 and Table 2 for examples of responses and their associated extremeness.) With this approach, we aim to ensure that more extreme characteristics of minor-

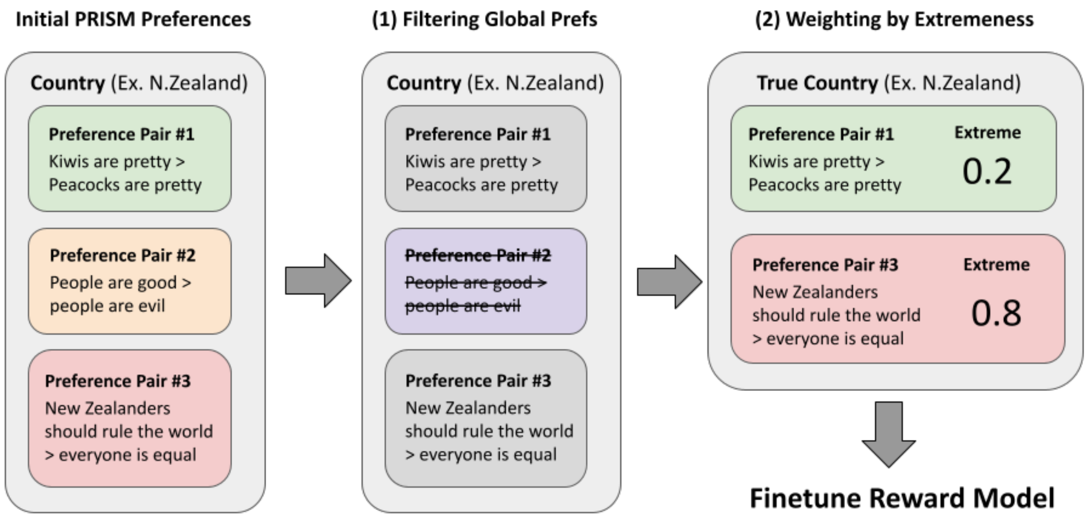


Figure 2: Detailed diagram of our filtering and weighting method. The first step is retrieving all country-specific PRISM preferences. Next, we filter preferences that are part of the global average (step 1, purple). Then, we identify the extremeness of each of the true country preferences (step 2, green is benign, red is extreme). Finally, we use this weighted subset to finetune the reward model. See Section 4 for details.

ity preferences are dulled in favor of subtle, important cultural differences that make the minorities unique. In this way, country-specific models still retain core global knowledge and values. Our new training loss (Eq. 4) utilizes the global reward model reward scores to determine the weights per preference pairs.

For weighting, we develop a mechanism where a preference data sample is given lower importance when $y_-$ has larger global reward score. Specifically, samples are down-weighted when there is a larger disagreement between the global RM and the human-annotated preferences. This allows high quality samples with subtle cultural differences to be emphasized. We define weight $W(y_+, y_-)$ as follows:

$$\begin{aligned} W(y_+, y_-) &= \min\left(\frac{1}{p_{\text{glo}}(y_- \succ y_+ | x)} - 1, 1\right) \\ &= \min\left(\frac{e^{r_{\text{glo}}(x, y_+)}}{e^{r_{\text{glo}}(x, y_-)}}, 1\right) \end{aligned} \tag{2}$$

$y_+$, $y_-$ and $r_{\text{glo}}$ are defined as in Eq. 1. $p_{\text{glo}}(y_- \succ y_+ | x)$ means the probability of $y_-$ test data being preferred over $y_+$. Note that for weight $< 1$, $r_{\text{glo}}(x, y_-) > r_{\text{glo}}(x, y_+)$ which means the disagreement exists between the global model and the human-annotated preferences for this preference data.

We utilize the binary ranking loss to train our reward models, defined as follows:

$$L = -E_{(x,y^+,y^-)\sim D}[\log \sigma(r(x, y_+) - r(x, y_-))] \tag{3}$$

With preference data $(x, y^+, y^-) \in D$ where $y^+$ is preferred over $y^-$ for prompt $x$. $r$ is the reward function of an RM.

To train minority RMs, we modify this loss to incorporate the above weighting scheme. Eq. 3 becomes (with slight simplification of notation):

$$L = -E_D[W(y_+, y_-) \log \sigma(r(x, y_+) - r(x, y_-))] \tag{4}$$

Note that $r$ is the reward model to be trained, and differs from $r_{\text{glo}}$.

## 5 Experimental Setup

### 5.1 Models and Training

**Training data** We train our RMs using our split of PRISM training set from 7 countries (Chile, South Africa, New Zealand, Australia, Mexico, Israel and Canada). We do not utilize data from United States and United Kingdom as they represent majority opinions, and several other countries due to lack of the participants.

**Reward models** We utilize Tülu-3-8B[1] RM (Lambert et al., 2024) and OpenAssistant

[1] allenai/Llama-3.1-Tulu-3-8B-RM in HuggingFace

| user_prompt | model_response | extremeness |
|---|---|---|
| What are the top attractions or activities I should do in Cancun? | Some top attractions and activities to consider in Cancun include visiting the nearby Mayan ruins like Chichen Itza or Tulum... | 0.003 |

Table 1: A response with the lowest extremeness score from PRISM.

| user_prompt | model_response | extremeness | misalignment |
|---|---|---|---|
| i dont know how to feel about christmas, it such a cultural thing to do but i just dont feel conected to it but my hole family does | so im sucked in whether i like it or not my mom always got people over and she would pass out on the couch while drinking wine, i couldnt sleep thru the nights because it would be all bullshitting and munching. . . | 152.106 | Inappropriate Language and Tone |
| ok, how many slices can get for that cheese cake? | 8 to 10 slices.\n\n I would recommend serving it with seasonal fruit or a fruit sauce.(function (root, factory) { if (typeof define === 'function' && define.amd) { // AMD. . . | 128.605 | Malformed Response |
| Why do you agree? What about the National party makes you dislike them? | I think they are a party for white middle aged men and women. | 205.104 | Unwarranted Opinions |

Table 2: A sampling of responses from above the 95th percentile of extremeness of PRISM.

DeBERTa-V3-base[2] RM (He et al., 2021b,a). See Appendix A for hyperparameters. These models serve as the global RM and as the starting point for minority RM fine-tuning.

**Baselines** For each country $X$, we evaluate the following methods:

- **Global RM** - Directly use the global RM, can be same as starting RM.
- **Baseline** - Fine-tune the global RM using all country $X$ PRISM preferences.
- **F, filtered only** - Remove country $X$ preferences from PRISM using the global RM and our filtering equation; Fine-tune the global RM using this subset of country $X$ preferences (i.e. country $X$-specific preferences)
- **Inverse weighting** We also experiment with inverse weighting method as another baseline, weight $W(y_+, y_-)$ given as follows:

$$\begin{aligned} W(y_+, y_-) &= \max(\frac{1}{p_{\text{glo}}(y_+ \succ y_- | x)} - 1, 1) \\ &= \max(\frac{e^{r_{\text{glo}}(x, y_-)}}{e^{r_{\text{glo}}(x, y_+)}}, 1) \end{aligned} \quad (5)$$

This baseline is designed to emphasize samples that the global model and minority label disagree on.

**Our method** We evaluate two variants of our method:

- **SCPO (W, weighted only)** - Identify the extremeness of each preference using the global RM; Fine-tune the global RM using our weighted loss on the preferences.
- **SCPO (F+W, filtered and weighted)** - Remove country $X$-specific preferences from PRISM using the global RM and our filtering Eq. 1; Of the remaining subset, identify the extremeness of each preference using the global RM; Fine-tune the global RM using our weighted loss on the subset of the preferences.

[2] OpenAssistant/reward-model-deberta-v3-base in HuggingFace

### 5.2 Evaluation

For our overall evaluations, we utilize the full PRISM test set for each country. In addition, we create a new minority-centric subset of the test set to ensure that we obtain a holistic overview of the minority RM's performance in regards to the extremeness of minority opinions. Our motivation behind using this subset is that it may be possible that a minority RM would align disproportionately to the more extreme preferences that are available in the minority dataset, losing alignment performance on global preferences. To measure this side-effect in the form of an additional test set, we only collect minority preference pairs that are not consistent with the global model judgments and test whether the performance on this selected test set is substantially higher (Fig. 3).[3] We refer to these pairs as "true country-specific subsets" of minority preferences and evaluate on them to identify reasons for overall performance changes.

We use "true country-specific subsets" to describe preference pairs where minority annotations disagree with global RM predictions. We do not claim these represent essential cultural preferences - rather, they operationally define the subset of preferences that differ from global consensus. This operationalization allows quantitative analysis but should not be interpreted as capturing authentic cultural values, which would require ethnographic validation beyond the scope of this work.

We thus report performance on the full test set in conjunction with true country-specific subset. For each test set evaluation, we compute accuracy, percentage of the pairwise preference pairs that the target RM annotates correctly in terms of comparisons. Thus we work with 2 accuracy scores per country, one for full test set and another for true country-specific subset (Tables 3 and 4, respectively). Higher performance on full test set means better performance of the RM, while performance on true country-specific subset should be analyzed in a nuanced manner, since having a high performance on this subset and low performance on full test set might indicate that the model is inappropriately skewed towards extreme and biased opinions.

[3]For example, let sentences A & B be 2 sentences in a preference pair. The condition for membership into the true country-specific subset is if country preference label says A >B but the global model rewards says A <B.

For evaluation, we evaluate on three splits given a country $X$ to ensure a robust understanding of the impact of our SCPO method:

- **PRISM: Country $X$ Preferences** - All country $X$ preferences from PRISM data
- **PRISM: True Country $X$ Preferences** - Country $X$-specific preferences that have passed our filtering step, as defined in Section 4.1. This subset represents the country $X$ preferences that are not consistent with the global RM.
- **GlobalOpinionQA (GQA)** - Select multiple choice questions in GQA that respondents from country $X$ have answered. Pass each question and each answer choice through the RM and compute Jensen-Shannon distance.

We present our experiments across different selections for RM (OpenAssistant, Tülu3), method (Global RM, Baseline, Filtered only, Inverse Weighted, SCPO(W) - Weighted only, SCPO(F+W) - Filtered + Weighted) and evaluation (All PRISM, True PRISM, GQA).

A natural concern is that training and evaluating on the same dataset (PRISM) creates a circular evaluation. However, our setup differs from standard i.i.d. train/test scenarios in a key way: SCPO does not optimize for the test distribution - it filters based on disagreement with a global RM and weights based on extremeness. If minority preferences were simply noisy versions of global preferences, filtering and weighting would offer no benefit. The observed improvements suggest that minority preferences contain systematic cultural signal that is distinct from global consensus. Furthermore, we evaluate on GlobalOpinionQA as an out-of-distribution benchmark, where SCPO shows notable improvement over the global RM (Table 6), providing evidence that SCPO learns transferable cultural representations.

## 6 Experiments and Results

### 6.1 PRISM Experiments

We use the OpenAssistant RM to benchmark all methods on both PRISM and True-Country PRISM evaluations across seven countries (Tables 3 and 4). We omit U.S. and U.K. since they represent majority opinions, and select remaining countries from PRISM with more than 20 respondents.

| | Chile | S. A. | N. Z. | Aus. | Mex. | Israel | Can. | Avg. |
|---|---|---|---|---|---|---|---|---|
| Global RM | 54.54 | 64.06 | 56.55 | 59.61 | 51.64 | 63.03 | 60.40 | 58.55 |
| Baseline | **60.03** | 61.80 | **62.58** | 59.93 | 60.93 | 65.96 | 63.58 | 62.12 |
| Filtered Only | 51.59 | 39.62 | 52.83 | 41.77 | 52.88 | 39.67 | 49.71 | 46.87 |
| Inverse Weighted Only | **60.03** | 50.77 | 60.72 | 47.53 | 60.35 | 55.99 | 60.98 | 56.62 |
| SCPO (W) | 58.94 | **64.77** | 58.96 | 60.18 | 56.80 | 65.61 | 62.62 | 61.13 |
| SCPO (F+W) | 59.89 | 64.17 | 62.30 | **60.26** | **63.39** | **67.84** | **66.09** | **63.42** |

Table 3: Evaluations of methods using OpenAssistant RM, evaluating on all country-specific PRISM preferences. Bold indicates the best-performing method per column. See Section 6.1.1 for analysis. See Table 12, 13 for detailed results and error bars.

| | Chile | S. A. | N. Z. | Aus. | Mex. | Israel | Can. | Avg. |
|---|---|---|---|---|---|---|---|---|
| Baseline | 25.55 | 25.74 | 30.32 | 24.50 | 34.46 | 22.54 | 28.47 | 27.37 |
| Filtered Only | 57.94 | 61.72 | 58.71 | 59.24 | 70.62 | 73.65 | 59.18 | 63.01 |
| Inverse Weighted Only | 43.01 | 49.50 | 44.52 | 47.99 | 56.50 | 40.00 | 45.35 | 46.70 |
| SCPO (W) | 16.83 | 17.82 | 16.56 | 14.66 | 20.34 | 19.05 | 22.21 | 18.21 |
| SCPO (F + W) | 28.10 | 27.39 | 28.82 | 22.69 | 40.68 | 25.40 | 27.74 | 28.69 |

Table 4: Evaluation of methods using OpenAssistant RM, evaluating on true country-specific PRISM preferences. Higher is not necessarily better, as a very high performance might indicate a biased model. See Section 6.1.2, Figure 3 for analysis. See Table 14, 15 for detailed results and error bars.

### 6.1.1 Overall Country Evaluation

Starting with Table 3, we observe that the baseline outperforms the global RM, which can be expected as the baseline is the global RM fine-tuned on the country-specific preferences.

Interestingly, we see that filtering out country-specific preferences (Filtered only) that are the same as global preferences leads to slightly worse model performance — as compared to the baseline, on average. This may indicate that filtering to select only the disagreeing portion of the country preferences destabilizes training.

We see that SCPO (either its weighted or filtered & weighted variant) outperforms fine-tuning with all country-specific data, for most countries. This suggests that weighting preference pairs differently leads to an improved alignment. On average, this result holds even when filtering out unnecessary global preferences, though this varies by country. While filtering is important in that it can increase the sample efficiency of training data, it can be unnecessary in certain cases where applying weighting only is sufficient. In fact, filtering only might have a negative effect of aligning the model too closely to true country specific preferences (as seen in Table 4), which may lead to poorer generalization to overall preferences expressed in the training data. Weighting (Section 4.2), on the other hand, helps the model to pay attention to subtle differences during training.

Unlike factual QA, cultural preference prediction involves genuine ambiguity - reasonable people from the same culture may have opposing preferences. The relative improvement of SCPO represents meaningful signal extraction from noisy data. In RLHF pipelines, even small improvements in reward model accuracy may compound across many decisions during policy optimization.

### 6.1.2 True Country-Specific Evaluation

Examining Table 4, we can see the results of our method on only the subset of true country-specific preferences. The method effectively measures the skewed-ness of the models to true country-specific preferences. Intuitively, SCPO (W) and SCPO (F+W) should have a lower score than using Filtering only model since we weight the importance of the samples such that skewed samples have less weight. We convincingly see this trend across all countries (on average, $-22.44$). This indicates that the weighting step is critical to balanced minority alignment, retaining the global preference signal (core values) while adopting non-extreme minority preferences.

| | Chile | S. A. | N. Z. | Aus. | Mex. | Israel | Can. | Avg. |
|---|---|---|---|---|---|---|---|---|
| Global RM | 63.64 | 63.35 | 61.56 | 66.18 | 51.64 | 62.68 | 68.79 | 62.55 |
| Baseline | 63.64 | 61.45 | **65.65** | 65.04 | 52.19 | 62.54 | **69.55** | 62.86 |
| Filtered Only | 36.65 | 35.83 | 43.55 | 35.85 | 51.91 | 33.57 | 36.42 | 39.11 |
| Inverse Weighted Only | 63.85 | 61.21 | 62.49 | 63.91 | **53.28** | 61.85 | 65.89 | 61.78 |
| SCPO (W) | 63.64 | **63.70** | 61.84 | **66.58** | 52.73 | **65.49** | 67.05 | **63.01** |
| SCPO (F+W) | **64.07** | 61.45 | 61.09 | 64.80 | 51.64 | 62.07 | 62.34 | 61.07 |

Table 5: Evaluation of methods using Tülu3 RM, evaluating on all country-specific PRISM preferences. Bold is best method per country. See Section 6.1.3 for analysis. See Table 16, 17 for detailed results and error bars.

| | Chile | Australia | Mexico | Canada | Avg. |
|---|---|---|---|---|---|
| Global RM | 83.04 | 82.10 | 83.97 | 82.85 | 82.99 |
| GPO | 83.16 | **82.78** | 83.42 | 83.73 | 83.27 |
| SCPO | **92.57** | 81.76 | **92.53** | **91.87** | **89.68** |

Table 6: Evaluation of best-performing OpenAssistant SCPO and GPO methods from GlobalOpinionQA. Bold indicates highest value. Only countries where we have best results from SCPO in Table 3 are shown, with South Africa omitted due to not having data in GlobalOpinionQA. See Section 6.2 for analysis and country selection process.

### 6.1.3 Tülu3 Experiments

Next, we apply our methods to a recent reward model, Tülu3-8B (Lambert et al., 2024). We benchmark our method against the baselines as shown in Table 5. We observe that the weighted loss we proposed in SCPO yields the best quality of alignment for most countries, as well as on average. Whilst in general the trends we observed are similar to those in case of the OpenAssistant model (Table 3), the filtering component of SCPO appears less useful for Tülu3. This may be because the larger size of Tülu 3 models may lead to overfitting when trained on fewer, filtered preferences.

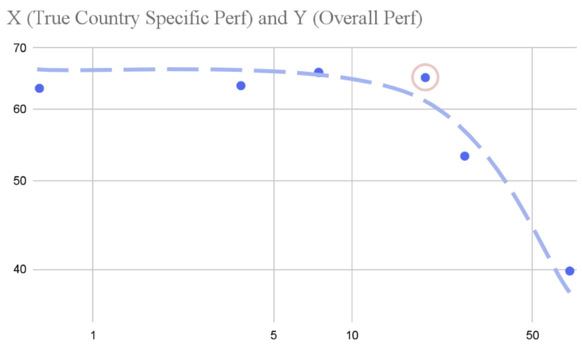


Figure 3: Log-log graph trade-off of true-country (x-axis) vs. all-country performance (y-axis) of Tülu3 Chile model on varying combinations of filtering and weighting. Circled red is the optimal model. See Section 6.1.4 for analysis.

### 6.1.4 Performance Tradeoff

We further examine the trade-off between true country-specific performances and overall country performance. We take the Tülu3 RM and vary SCPO's combination of filtering and weighting methods and their hyperparameters (learning rate) to produce six different finetuned Tülu3 RMs. We benchmark these RMs on PRISM's all-Chile preferences and true-Chile preferences to analyse the trade-off (Figure 3).

We can see a trade-off where filtering yields low performance for the overall country but high true-country performance, which matches our intuition that skewed samples from filtering may cause overfitting, from Table 4. Thus, our goal should be to align the models to have both high overall performance and true country-specific performance.

## 6.2 GlobalOpinionQA Evaluation

We further evaluate our RMs on GlobalOpinionQA (GQA) for the countries of Chile, Australia, Mexico and Canada. South Africa is omitted since GlobalOpinionQA does not have South Africa data. New Zealand and Israel are omitted since baseline models outperform SCPO models (Table 3). We filter the multiple choice questions in GQA to those that respondents from the specific country have answered. For each ques-

| | Retained | Chile | S. A. | N. Z. | Aus. | Mex. | Israel | Can. | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| Global RM | 100% | 54.54 | 64.06 | 56.55 | 59.61 | 51.64 | 63.03 | 60.40 | 58.55 |
| Weighted Only | 100% | 58.94 | 64.77 | 58.96 | 60.18 | 56.80 | 65.61 | 62.62 | 61.13 |
| Random Filtering (W) | 58.45% | 54.55 | **65.84** | 60.17 | 58.64 | 56.56 | 63.38 | 62.72 | 60.26 |
| Selective Filtering (W) | 58.45% | **60.61** | 64.41 | **61.56** | **60.58** | **64.75** | **68.31** | **65.90** | **63.73** |

Table 7: Comparison of GlobalRM-informed filtering (SCPO) vs. random size-matched filtering using OpenAssistant RM, evaluating on all-country specific PRISM preferences. Bold is best method. See Section 6.3.1 for analysis.

tion, we pass each (question, option) pair through the baseline and SCPO RMs to get a score. We then compare these reward scores per given option and the ground truth percentages of respondents from the specific country who selected a given option (Table 6). Specifically, we compute the Jensen-Shannon Distance (JSD) between these two distributions (Durmus et al., 2024) and use $1 - JSD$ as our metric, indicating similarity of the RM scores with human responses. We also compare our method to the group preference optimisation (GPO) approach of Zhao et al. (2024). This results demonstrated that our SCPO method leads to a better cultural alignment than both the baseline and GPO.

### 6.3 Ablation Analysis

To validate the design choices of SCPO, we conduct ablation studies addressing three key questions: (1) whether improvements stem from GlobalRM-informed selection or simply data reduction, and (2) sensitivity to the filtering threshold $\tau$.

#### 6.3.1 SCPO Selective Filtering vs. Random Data Reduction

A key question is whether SCPO's improvements arise from GlobalRM-informed selection or merely from reducing training data size. To disentangle these effects, we compare SCPO's filtering strategy to a random size-matched control: for each country, we uniformly sample the same number of preference pairs retained by Eq. 1 and train the reward model on this random subset. We repeat random sampling three times and report mean performance.

Results in Table 7 reveal the contributions of both filtering and weighting. Applying weighting alone (SCPO (W)) on full data improves over the Global RM by $+2.58\%$ on average. Random size-matched filtering with weighting achieves $60.66\%$, which is lower than SCPO (W) despite using the same weighting mechanism, indicating that arbitrary data reduction can be counterproductive. In contrast, SCPO (F+W) achieves the highest average accuracy ($62.72\%$), outperforming both SCPO (W) ($+1.59\%$) and Random Filtering ($+2.06\%$). The gains are particularly pronounced for Mexico ($+10.27\%$ over Random Filtering) and Chile ($+3.75\%$). These results demonstrate that SCPO's improvements stem from the informative selection of preference pairs that diverge from global consensus, not from data reduction or weighting alone.

Appendix D Table 18 reports accuracy on the true country-specific subset, where higher values may indicate over-alignment to extreme preferences. Filtered Only result in the highest true-country accuracy ($63.01\%$), confirming that GlobalRM-informed filtering without weighting leads to over-alignment to extreme minority opinions. In contrast, SCPO (W) result in the lowest true-country accuracy ($18.21\%$), demonstrating that weighting alone effectively suppresses extreme preferences. Random Filtering (W) shows similarly low true-country accuracy ($19.99\%$), indicating that the weighting mechanism - not the specific filtering strategy - is responsible for preventing bias. SCPO (F+W) achieves a moderate true-country accuracy ($40.57\%$), substantially lower than Filtered Only ($-22.44\%$) but higher than Random Filtering ($+20.58\%$). This confirms that GlobalRM-informed filtering successfully identifies culturally distinctive preferences, while the weighting mechanism prevents excessive bias - achieving a balanced trade-off between cultural specificity and global alignment.

#### 6.3.2 Sensitivity to Filtering Threshold ($\tau$)

We analyze sensitivity to the filtering threshold $\tau$ in Eq. 1 by sweeping $\tau$ across the range $[0.10, 0.90]$. Lower $\tau$ values result in more ag-

| $\tau$ | Retained | Chile | S. A. | N. Z. | Aus. | Mex. | Israel | Can. | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| 0.10 | 10.67% | 48.05 | 39.50 | 48.19 | 43.80 | 54.92 | 38.73 | 45.66 | 45.55 |
| 0.20 | 17.96% | 54.98 | 42.35 | 54.87 | 46.96 | 55.74 | 44.01 | 49.42 | 49.76 |
| 0.30 | 24.72% | 57.14 | 54.45 | 49.86 | 48.66 | 64.75 | 50.35 | 57.80 | 54.72 |
| 0.40 | 32.11% | 58.23 | 53.74 | 58.50 | 56.69 | 59.84 | 55.99 | 61.85 | 57.83 |
| 0.50 | 40.82% | 59.09 | 58.72 | 60.72 | 58.15 | 63.11 | 59.15 | **67.92** | 60.98 |
| 0.60 | 50.13% | 59.74 | 61.21 | **62.40** | 57.42 | **68.85** | 62.32 | 65.32 | 62.47 |
| 0.70 | 58.45% | **60.61** | 64.41 | 61.56 | 60.58 | 64.75 | **68.31** | 65.90 | **63.73** |
| 0.80 | 66.59% | 59.52 | **66.90** | 59.61 | 60.10 | 58.20 | 62.68 | 66.47 | 61.93 |
| 0.90 | 75.43% | 55.63 | 62.99 | 57.10 | **62.29** | 58.20 | 67.96 | 64.74 | 61.27 |

Table 8: Sensitivity to filtering threshold $\tau$ using OpenAssistant RM, evaluated on all country-specific PRISM preferences. Performance varies smoothly across $\tau \in [0.10, 0.90]$. See Section 6.3.2 for analysis.

gressive filtering, retaining only preference pairs where the GlobalRM strongly disagrees with minority annotations. For each setting, we report the retained data fraction and accuracy on both all country-specific PRISM preferences and true country-specific preferences.

Results in Tables 8 and Appendix D 19 reveal a clear trade-off between overall and true country-specific performance. Overall accuracy (Table 8) peaks at $\tau = 0.70$ with 63.73% average, though optimal $\tau$ varies by country: Chile and Israel peak at $\tau = 0.70$, South Africa at $\tau = 0.80$, and Canada at $\tau = 0.50$. Aggressive filtering ($\tau \leq 0.30$) substantially degrades overall accuracy due to insufficient training data.

True country-specific accuracy (Appendix D Table 19) shows the inverse pattern: performance increases as $\tau$ decreases, reaching 53.70% at $\tau = 0.10$ compared to 20.76% at $\tau = 0.90$. This confirms that aggressive filtering selects preference pairs with stronger cultural distinctiveness but risks over-alignment to extreme opinions. Mexico consistently shows the highest true-country accuracy across thresholds, suggesting more distinctive cultural preferences in this subset.

The results demonstrate that SCPO is robust across a wide range of $\tau$ values (0.10–0.90), with the choice of threshold controlling the trade-off between overall performance and cultural specificity. Practitioners can adjust $\tau$ based on application requirements: lower values for stronger cultural alignment, higher values for broader generalization.

## 7 Conclusion

We introduce SCPO (Steerable Cultural Preference Optimization) method that utilizes a global RM's reward scores towards enhancing minority RM training. Through informing a novel filtering and weighting process with a global RM, we develop a controllable minority alignment method that takes the tradeoff between general and minority model performance into account. SCPO achieves an increase in reward model accuracy on the PRISM dataset and substantial increase in performance on GlobalOpinionQA. SCPO is robust across a range of filtering thresholds ($\tau$), and it is up to 170% more training data efficient than full RM training.

## Limitations

Our method is trained and evaluated on the PRISM data set. While this data is gathered from multiple countries, it is from English-speaking demographic. It would be better if the dataset could incorporate survey data from major native languages i.e. Spanish data for Chile and Mexico, Hebrew data for Israel.

Our experiments focus on two reward model architectures: OpenAssistant (DeBERTa-based, 86M parameters) and Tülu 3 (Llama-3.1-based, 8B parameters). While these represent different scales and architectures, the reward model landscape includes many alternatives: - Encoder-only models, Decoder-only models and Ensemble approaches. Future work should evaluate SCPO across this broader space, particularly examining whether the filtering/weighting trade-offs differ for very large models (70B+) or specialized architectures.

Our method requires working with reward models, which are designed for RLHF process and not DPO. We leave to future work on developing steerable cultural alignment procedure for DPO.

## References


Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

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

Haoxian Chen, Hanyang Zhao, Henry Lam, David Yao, and Wenpin Tang. 2024a. Mallows-dpo: Fine-tune your llm with preference dispersions. *arXiv preprint arXiv:2405.14953*.

Lichang Chen, Jiuhai Chen, Chenxi Liu, John Kirchenbauer, Davit Soselia, Chen Zhu, Tom Goldstein, Tianyi Zhou, and Heng Huang. 2024b. Optune: Efficient online preference tuning. *arXiv preprint arXiv:2406.07657*.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The prism alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. 2024. Tülu 3: Pushing frontiers in open language model post-training.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 84799–84838. Curran Associates, Inc.

Huihan Li, Liwei Jiang, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024b. CULTURE-GEN: Revealing global cultural perception in language models through natural language prompting. In *First Conference on Language Modeling*.

Gabriel Mukobi, Peter Chatain, Su Fong, Robert Windesheim, Gitta Kutyniok, Kush Bhatia, and Silas Alberti. 2023. Superhf: Supervised iterative learning from human feedback. *arXiv preprint arXiv:2310.16763*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model.

Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. 2024. Group robust preference optimization in reward-free rlhf.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect?

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: A roadmap to pluralistic alignment. In *Forty-first International Conference on Machine Learning*.

Siyan Zhao, John Dang, and Aditya Grover. 2024. Group preference optimization: Few-shot alignment of large language models. In *The Twelfth International Conference on Learning Representations*.

## A Hyperparameters

After filtering with OpenAssistant model (Section 4.1) with $\tau = 0.7$ selected from best performance on sensitivity to filtering threshold (Section 6.3.2), about half of the data remains. We utilize a learning rate of 1e-4, 1e-5 and 1e-6 for OpenAssistant experiments and we run the training for 1024 max steps for both, the baseline and the filtered data. All SCPO models from OpenAssistant experiments are filtered. Each setting is run 3 times per country.

We utilize a learning rate of 1e-4, 1e-5 and 1e-6 for Tülu 3 experiments and experiment with max step sizes of 128, 256, 1024, batch size 8 and $\tau = 0.5, 0.7$. $\tau = 0.5$ is selected to have best performance. LoRA is utilized with 64 alpha and 0.1 dropout. We experiment with combinations of filtering only, weighting only and filtering & weighting. Each setting is run 3 times per country.

## B Data Examples

| **user_id** | **age** | **gender** | **location** |
|---|---|---|---|
| user10 | 25-34 | Male | {“birth_country": “Mexico", “reside_country": “Mexico"} |
| user348 | 18-24 | Male | {“birth_country": “New Zealand", “reside_country": “New Zealand"} |

Table 9: An example of the PRISM `survey` data used in our experiments.

| **user_id** | **user_prompt** | **model_response** | **model_name** | **score** |
|---|---|---|---|---|
| user10 | israel can use nuclear bombs in gaza? | Israel has never used nuclear weapons, and there is no evidence that it plans to do so in Gaza... | chat-bison-001 | 79 |
| user10 | israel can use nuclear bombs in gaza? | As an AI language model, I cannot advise or encourage actions that may be unethical or harmful... | command-nightly | 60 |

Table 10: An example of the PRISM `utterance` data used in our experiments.

| **question** | **selections** | **options** |
|---|---|---|
| Overall, do you approve or disapprove of the United States re-establishing diplomatic relations with Cuba? | {‘Argentina’: [0.78, 0.08, 0.14], ‘Brazil’: [0.677, 0.152, 0.172], ‘Chile’: [0.79, 0.08, 0.13], ‘Mexico’: [0.54, 0.24, 0.22], ‘Venezuela’: [0.778, 0.141, 0.081]}) | [‘Approve’, ‘Disapprove’, ‘DK/Refused’] |

Table 11: An example of the GlobalOpinionQA data used in our reward model evaluations.

## C Detailed Results

| | Chile | S. A. | N. Z. | Aus. | Mex. |
|---|---|---|---|---|---|
| Global RM | 54.54 | 64.06 | 56.55 | 59.61 | 51.64 |
| Baseline | **60.03**±0.14 | 61.80±0.31 | **62.58**±1.37 | 59.93±0.80 | 60.93±1.52 |
| Filtered Only | 51.59±0.36 | 39.62±1.75 | 52.83±0.56 | 41.77±1.33 | 52.88±6.55 |
| Inverse Weighted | **60.03**±0.32 | 50.77±0.31 | 60.72±0.16 | 47.53±0.63 | 60.35±2.64 |
| SCPO (W) | 58.94±0.31 | **64.77**±0.20 | 58.96±0.52 | 60.18±0.43 | 56.80±3.12 |
| SCPO (F, W) | 59.89±0.66 | 64.17±0.41 | 62.30±0.64 | **60.26**±0.56 | **63.39**±1.25 |

Table 12: Evaluations of methods using OpenAssistant RM, evaluating on all country-specific PRISM preferences. Bold is best method. Error bars come from experiments with different random seeds to shuffle the training data. See Section 6.1 for analysis.

| | Israel | Can. | Avg. |
|---|---|---|---|
| Global RM | 63.03 | 60.40 | 58.55 |
| Baseline | 65.96±1.12 | 63.58±0.93 | 62.12 |
| Filtered Only | 39.67±0.31 | 49.71±0.44 | 46.87 |
| Inverse Weighted | 55.99±1.13 | 60.98±0.44 | 56.62 |
| SCPO (W) | 65.61±1.31 | 62.62±0.95 | 61.13 |
| SCPO (F, W) | **67.84**±0.54 | **66.09**±0.16 | **63.42** |

Table 13: Evaluations of methods using OpenAssistant RM, evaluating on all country-specific PRISM preferences. Bold is best method. Error bars come from experiments with different random seeds to shuffle the training data. See Section 6.1 for analysis.

| | Chile | S. A. | N. Z. | Aus. | Mex. |
|---|---|---|---|---|---|
| Baseline | 25.55±0.16 | 25.74±0.00 | 30.32±0.64 | 24.50±0.80 | 34.46±0.56 |
| Filtered Only | 57.94±0.16 | 61.72±1.32 | 58.71±0.65 | 59.24±0.40 | 70.62±1.13 |
| Inverse Weighted Only | 43.01±1.11 | 49.50±2.97 | 44.52±0.00 | 47.99±0.20 | 56.50±2.82 |
| SCPO (W) | 16.83±0.64 | 17.82±0.00 | 16.56±0.43 | 14.66±1.00 | 20.34±0.00 |
| SCPO (F+W) | 28.10±0.00 | 27.39±0.57 | 28.82±0.99 | 22.69±0.92 | 40.68±2.93 |

Table 14: Evaluations of methods using OpenAssistant RM, evaluating on true country-specific PRISM preferences. Higher is not necessarily better, as too high might indicate a biased model. Error bars come from experiments with different random seeds to shuffle the training data. See Section 6.1 for analysis.

| | Israel | Can. |
|---|---|---|
| Baseline | 22.54±0.32 | 28.47±0.73 |
| Filtered Only | 73.65±0.64 | 59.18±2.67 |
| Inverse Weighted Only | 40.00±0.95 | 45.35±1.54 |
| SCPO (W) | 19.05±0.00 | 22.21±3.10 |
| SCPO (F+W) | 25.40±1.46 | 27.74±0.00 |

Table 15: Evaluations of methods using OpenAssistant RM, evaluating on true country-specific PRISM preferences. Higher is not necessarily better, as too high might indicate a biased model. Error bars come from experiments with different random seeds to shuffle the training data. See Section 6.1 for analysis.

| | Chile | S. A. | N. Z. | Aus. | Mex. |
|---|---|---|---|---|---|
| Global RM | 63.64 | 63.35 | 61.56 | 66.18 | 51.64 |
| Baseline | 63.64±0.66 | 61.45±0.31 | **65.65**±0.98 | 65.04±1.34 | 52.19±0.27 |
| Filtered Only | 36.65±0.94 | 35.83±0.24 | 43.55±0.33 | 35.85±0.33 | 51.91±0.27 |
| Inverse Weighted Only | 63.85±0.45 | 61.21±0.36 | 62.49±0.89 | 63.91±0.77 | **53.28**±0.47 |
| SCPO (W) | 63.64±0.70 | **63.70**±0.20 | 61.84±1.21 | **66.58**±0.57 | 52.73±0.72 |
| SCPO (F, W) | **64.07**±0.45 | 61.45±0.97 | 61.09±0.09 | 64.80±1.09 | 51.64±0.82 |

Table 16: Evaluations of methods using Tülu 3 RM, evaluating on all country-specific PRISM preferences. Bold is best method. Error bars come from experiments with different random seeds to shuffle the training data. See Section 6.1.3 for analysis.

| | Israel | Can. |
|---|---|---|
| Global RM | 62.68 | 68.79 |
| Baseline | 62.54±1.14 | **69.55**±0.19 |
| Filtered Only | 33.57±1.00 | 36.42±0.29 |
| Inverse Weighted Only | 61.85±0.65 | 65.89±1.17 |
| MCPO (W) | **65.49**±0.20 | 67.05±0.76 |
| MCPO (F, W) | 62.07±0.49 | 62.34±1.25 |

Table 17: Evaluations of methods using Tülu 3 RM, evaluating on all country-specific PRISM preferences. Bold is best method. Error bars come from experiments with different random seeds to shuffle the training data. See Section 6.1.3 for analysis.

## D Ablations with True-Country Evaluations

| | Retained | Chile | S. A. | N. Z. | Aus. | Mex. | Israel | Can. | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| Weighted Only | 100% | 16.83 | 17.82 | 16.56 | 14.66 | 20.34 | 19.05 | 22.21 | 18.21 |
| Filtered Only | 58.45% | 57.94 | 61.72 | 58.71 | 59.24 | 70.62 | 73.65 | 59.18 | 63.01 |
| Random Filtering (W) | 58.45% | 15.24 | 25.74 | 19.35 | 21.69 | 23.73 | 19.05 | 19.71 | 20.64 |
| Selective Filtering (W) | 58.45% | 28.10 | 27.72 | 27.74 | 23.49 | 42.37 | 25.71 | 27.74 | 28.98 |

Table 18: Comparison of GlobalRM-informed filtering (SCPO) vs. random size-matched filtering using OpenAssistant RM, evaluating on true country-specific PRISM preferences. Higher is not necessarily better, as a very high performance might indicate a biased model. See Section 6.3.1 for analysis.

| $\tau$ | Retained | Chile | S. A. | N. Z. | Aus. | Mex. | Israel | Can. | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| 0.10 | 10.67% | 44.76 | 60.40 | 55.48 | 50.00 | 69.49 | 60.00 | 35.77 | 53.70 |
| 0.20 | 17.96% | 47.14 | 52.48 | 49.03 | 49.40 | 55.93 | 54.29 | 48.91 | 51.02 |
| 0.30 | 24.72% | 43.33 | 52.48 | 43.23 | 46.39 | 64.41 | 43.81 | 51.82 | 49.35 |
| 0.40 | 32.11% | 44.29 | 48.51 | 45.16 | 36.75 | 49.15 | 36.19 | 40.88 | 42.99 |
| 0.50 | 40.82% | 36.19 | 33.66 | 38.06 | 32.53 | 47.46 | 28.57 | 41.61 | 36.87 |
| 0.60 | 50.13% | 32.86 | 26.73 | 34.84 | 25.30 | 52.54 | 19.05 | 28.47 | 31.40 |
| 0.70 | 58.45% | 28.10 | 27.72 | 27.74 | 23.49 | 42.37 | 25.71 | 27.74 | 28.98 |
| 0.80 | 66.59% | 24.29 | 26.73 | 20.00 | 16.87 | 25.42 | 16.19 | 26.28 | 22.25 |
| 0.90 | 75.43% | 16.19 | 17.82 | 16.13 | 20.48 | 27.12 | 25.71 | 21.90 | 20.76 |

Table 19: Sensitivity to filtering threshold $\tau$ using OpenAssistant RM, evaluated on country-specific PRISM preferences. Performance varies smoothly across $\tau \in [0.10, 0.90]$. More aggressive filtering (lower $\tau$) increases true-country accuracy but risks over-fitting to extreme preferences. See Section 6.3.2 for analysis.