# Implicit Cross-Lingual Rewarding
# for Efficient Multilingual Preference Alignment

**Wen Yang[1,2], Junhong Wu[1,2], Chen Wang[1,2], Chengqing Zong[1,2], Jiajun Zhang[1,2,3] [*]**

[1] School of Artificial Intelligence, University of Chinese Academy of Sciences
[2] Institute of Automation, Chinese Academy of Sciences
[3] Wuhan AI Research
{yangwen2023, wujunhong2021, wangchen2020}@ia.ac.cn
{cqzong, jjzhang}@nlpr.ia.ac.cn

## Abstract

Direct Preference Optimization (DPO) has become a prominent method for aligning Large Language Models (LLMs) with human preferences. While DPO has enabled significant progress in aligning English LLMs, multilingual preference alignment is hampered by data scarcity. To address this, we propose a novel approach that *captures* learned preferences from well-aligned English models by implicit rewards and *transfers* them to other languages through iterative training. Specifically, we derive an implicit reward model from the logits of an English DPO-aligned model and its corresponding reference model. This reward model is then leveraged to annotate preference relations in cross-lingual instruction-response pairs, using English instructions to evaluate multilingual responses. The annotated data is subsequently used for multilingual DPO fine-tuning, facilitating preference knowledge transfer from English to other languages. Fine-tuning Llama3 for two iterations resulted in a 12.72% average improvement in Win Rate and a 5.97% increase in Length Control Win Rate across all training languages on the X-AlpacaEval leaderboard. Our findings demonstrate that leveraging existing English-aligned models can enable efficient and effective multilingual preference alignment, significantly reducing the need for extensive multilingual preference data. The code is available at https://github.com/ZNLP/Implicit-Cross-Lingual-Rewarding.

## 1 Introduction

Direct alignment algorithms (DAAs), such as DPO (Rafailov et al., 2024b) and its variants (Azar et al., 2024; Ethayarajh et al., 2024; Meng et al., 2024), renowned for their simplicity, efficiency and stability than Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022),
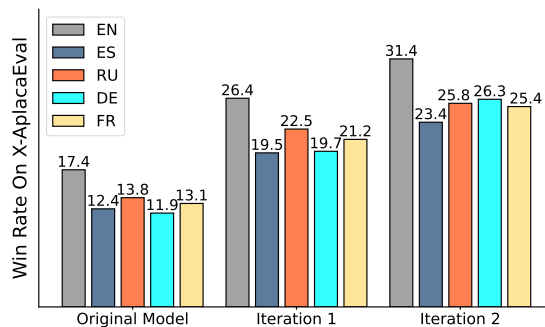


Figure 1: Iterative Preference Transfer and Improvement with **Implicit Cross-Lingual Rewarding** based on the English-aligned Llama3 model. Detailed results are shown in Table 2.

have emerged as valuable and widely adopted post-training techniques for aligning LLMs with human preferences. While English benefits from abundant high-quality preference datasets (Cui et al., 2024; Mukherjee et al., 2023) and has merged numerous DAAs-aligned models, multilingual preference alignment is challenged by data scarcity.

Existing approaches typically rely on expensive human annotation or advanced multilingual preference alignment models (Ahmadian et al., 2024; Dang et al., 2024) to annotate data for each language, thereby constructing off-policy multilingual preference datasets. However, this approach faces significant challenges due to the scarcity and cost of annotations, particularly for low-resource languages. Furthermore, translation-based methods either translate English preference data into other languages (Lai et al., 2023) or use translation to derive reward signals to construct multilingual preference data (She et al., 2024; Yang et al., 2024c). These methods can introduce artifacts and distort preference signals, hindering effective multilingual preference learning.

This work explores a novel perspective: *leveraging the preference knowledge embedded within existing English-aligned models to facilitate multilingual preference alignment.* Prior work (Chen et al.,

---

[*] Corresponding author

**Capture Learned Preferences in Model**  |  **Transfer Preferences to Multilingual**
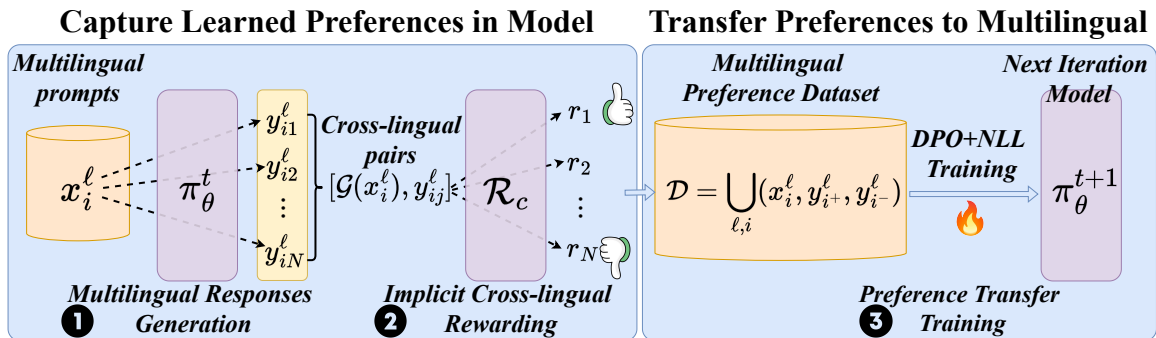
Figure 2: **Implicit Cross-Lingual Rewarding for Efficient Multilingual Preference Alignment.** Our method consists of three steps: (i) *Multilingual Responses Generation*: Sampling multilingual responses from parallel prompts with $\pi_\theta^t$, respectively. (ii) *Implicit Cross-lingual Rewarding*: Scoring these responses with cross-lingual instruction-response pairs, where instructions are mapped into English via $\mathcal{G}(x_i^\ell)$ (Eq. 8) and the pairs evaluated with the implicit cross-lingual reward $\mathcal{R}_c$ (Eq. 9) (iii) *Preference Transfer Training*: Preference pairs are constructed based on scores for DPO+NLL training, producing an improved model $\pi_\theta^{t+1}$. This process is repeated iteratively, gradually enhancing the model's multilingual preference alignment until optimization saturates.

2024a) has demonstrated that the implicit reward model, derived from the logits of a well-aligned English DPO model and its reference model, effectively captures preferences over English instructions. Building on this, we apply this implicit reward model to the multilingual setting, using it to label preference relations in cross-lingual instruction-response pairs (Li et al., 2024). This ensures that multilingual responses are evaluated based on their alignment with English instructions. We term *Implicit Cross-Lingual Rewarding*, which preserves reward signal fidelity by directly evaluating multilingual responses under English instructions, avoiding translation-induced distortions.

As shown in Figure 2, our approach involves three key steps: (1) Multilingual response generation: Starting from any multilingual model that is DPO-tuned on English preference data from an initial reference model. Responses are sampled by the model from multilingual prompts. (2) Implicit cross-lingual rewarding: Constructing cross-lingual instruction-response pairs by pairing English instructions with sampled multilingual responses. The implicit reward model then assigns preference scores to these responses, capturing the model's learned preference knowledge. (3) Preference Transfer Training: Our approach adopts iterative DPO similar to previous works (Yuan et al., 2024; Yang et al., 2024c), incorporating a negative log-likelihood (NLL) loss term to train on the multilingual preference data, thereby transferring preferences across languages.

Our experiments start with the existing English-aligned Llama 3 model, followed by two iterations of our training process. Results (Figure 1) demon-

strate that our approach not only transfers preference knowledge from English to other languages but also iteratively improves English alignment through implicit reward. This suggests that each iteration inherently facilitates both preference transfer and refinement within the multilingual LLM. Notably, experiments with other DAAs-aligned base models and lower-resource languages confirm the broad applicability of implicit cross-lingual rewarding, establishing it as an efficient and robust strategy for enhancing multilingual preference alignment for any English-aligned model.

## 2 Preliminaries

This section introduces two prominent methods in preference optimization, Reinforcement Learning with Human Feedback (RLHF) and Direct Preference Optimization (DPO), and derives the implicit rewards of the DPO-tuned model.

In preference optimization, the preference data typically takes the pairwise form, denoted as $\mathcal{D} = \{(x, y_w, y_l)\}$. Each prompt $x$ is paired with two possible responses, $y_w$ and $y_l$, where $y_w$ is designated as the preferred response by human evaluators.

### 2.1 Reinforcement Learning From Human Feedback

RLHF uses human feedback to adjust a model's behavior, typically by incorporating a reward model. Since directly modeling pairwise preferences between $y_w$ and $y_l$ is difficult, a common approach defines a reward function $r(x, y)$, from which preferences are inferred, often using the Bradley-Terry model (Bradley and Terry, 1952) to represent such

preferences.

$$p(y_w \succ y_l|x) = \frac{\exp(r(x, y_w))}{\exp(r(x, y_w)) + \exp(r(x, y_l))} \quad (1)$$

From this formulation, RLHF first trains a parameterized reward model $r_\phi(x, y)$ using maximum likelihood:

$$\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[\log \sigma(r_\phi(x, y_w) - r_\phi(x, y_l))\right] \quad (2)$$

Where $\sigma$ is the logistic function, then the objective of RLHF is to optimize the policy model $\pi_\theta$ to maximize the expected value of the reward function. Given the pre-trained RM $r_\phi(x, y)$ and a reference model $\pi_{ref}$ (typically an SFT model), the objective is to find a new model $\pi_\theta$ by maximizing the following expression.

$$\max_{\pi_\theta} \left\{ \mathbb{E}_{\mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})}[r(\mathbf{x}, \mathbf{y})] - \beta \log \frac{\pi_\theta(\mathbf{y}|\mathbf{x})}{\pi_{ref}(\mathbf{y}|\mathbf{x})} \right\} \quad (3)$$

Where KL divergence (Kullback and Leibler, 1951) from the reference policy $\pi_{ref}$ is usually incorporated as a regularization to prevent the reward over-optimization of $\pi_\theta$, $\beta$ controls the deviation from the base reference policy. The objective is then optimized using the RL algorithm, such as Proximal Policy Optimization (PPO) (Schulman et al., 2017).

## 2.2 Direct Preference Optimization

Unlike RLHF, which learns a reward model before optimizing it via reinforcement learning, Direct Preference Optimization (DPO) leverages a reward model parameterization that allows for closed-form extraction of the optimal policy, eliminating the RL training loop. DPO's key insight is to directly model pairwise preferences. Specifically, DPO models the probability of preferring response $y_w$ over response $y_l$ given prompt $x$ as:

$$p_\theta(y_w \succ y_l|x) = \sigma \left( \beta \log \frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)} \right) \quad (4)$$

Where $\sigma$ is the sigmoid function. DPO then directly trains the optimal model on human feedback data $\mathcal{D}$ by maximizing the likelihood of these pairwise preferences using the following objective:

$$\mathcal{L}(\pi_\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[\log p_\theta(y_w \succ y_l|x)\right] \quad (5)$$

**Implicit Reward in DPO-tuned Model** Thus, DPO directly implicitly learns the underlying reward function without a separate reward model training stage. The reward is parameterized in terms of the corresponding optimal policy $\pi_\theta$ and a reference policy $\pi_{ref}$:

$$r(x, y) = \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)} \quad (6)$$

# 3 Implicit Cross-Lingual Rewarding For Efficient Multilingual Alignment

Our approach leverages an existing English preference-aligned model and multilingual training prompts to iteratively improve preference alignment across all languages without external annotations. By exploiting the model's English preference alignment capabilities, we use implicit cross-lingual rewards to progressively enhance multilingual alignment. For illustrative purposes, we begin with DPO in our approach and then extend to other DAA in our experiments. The outline is shown in Figure 2, each iteration involves (1) sampling multilingual responses, (2) scoring these responses with implicit cross-lingual rewards, and (3) constructing multilingual preference pairs for DPO training.

**Initialization** Given any multilingual LLM $\pi_\theta^0$, that is DPO-tuned on English preference data from an initial reference model $\pi_{ref}$, and a set of parallel multilingual instructions $\mathcal{X}$, where $\mathcal{X}$ consists of English and other language instructions $(x^{en}, x^{es}, \ldots, x^{ru})$. After $T$ rounds training, the model is represented as $\pi_\theta^1, \pi_\theta^2, \ldots, \pi_\theta^T$.

**Multilingual Responses Generation** For each round $t \in \{1, 2, \ldots\}$, given input $x_i^\ell$, we sample $N$ responses, $y_{1\ldots N}^\ell$ from the model $\pi_\theta^t$, where $\pi_\theta^t$ is the latest policy model. Note that $\ell$ refers to any language supported by the model.

$$y_{1\ldots N}^\ell \sim \pi_\theta^t(x_i^\ell) \quad \text{for all } x_i^\ell \in \mathcal{X} \quad (7)$$

**Implicit Cross-Lingual Rewarding** For any LLM $\pi_\theta$ that has undergone direct alignment optimization in English, the resulting model implicitly embodies a reward model. The implicit reward model, denoted as $r(x, y)$, can be expressed in terms of $\pi_\theta$ and its reference model $\pi_{ref}$, as shown in Eq. (6).

To leverage the learned preference in $\pi_\theta$, we introduce a cross-lingual reward mechanism to effectively annotate multilingual preference data using implicit rewards. For responses generated from

prompts in other languages, we create cross-lingual instruction-response pairs using parallel English prompts and leverage $r(x, y)$ to score these pairs.

Specifically, we define a mapping function $\mathcal{G} : \mathcal{I}^\ell \to \mathcal{I}^{en}$, where $\mathcal{I}^\ell$ represents the space of instructions in language $\ell$, and $\mathcal{I}^{en}$ represents the space of English instructions. Given an instruction $x_i^\ell$ in language $\ell$, we construct its corresponding English instruction $\mathcal{G}(x_i^\ell)$, which is then used for reward scoring.

$$\mathcal{G}(x_i^\ell) = \begin{cases} x_i^{en} & \text{if } \ell = \text{en}, \\ \mathrm{P}(\ell) + x_i^{en} & \text{if } \ell \neq \text{en}. \end{cases} \quad (8)$$

> **Cross-lingual Instruction Prefix P$(\ell)$**
>
> Please answer the following instruction using only $\ell$ unless explicitly instructed to respond in a different language.

In this formalization, when the target language $\ell$ is English ($\ell = \text{en}$), the function returns the original instruction $x_i^{en}$. When the target language $\ell$ is not English ($\ell \neq \text{en}$), the function prepends a cross-lingual instruction prefix $\mathrm{P}(\ell)$, to the parallel English instruction $x_i^{en}$. This prefix $\mathrm{P}(\ell)$ incorporates a language constraint, ensuring that the resulting instruction $\mathcal{G}(x_i^\ell)$ is semantically aligned with the target language $\ell$ and compatible with the reward model.

To mitigate length exploitation (Park et al., 2024), a phenomenon observed in preference learning, we incorporate a length penalty used in RLHFlow (Dong et al., 2024). The cross-lingual reward $\mathcal{R}_c$ is then calculated as:

$$\mathcal{R}_c = \beta \log \frac{\pi_\theta(y \mid \mathcal{G}(x_i^\ell))}{\pi_{ref}(y \mid \mathcal{G}(x_i^\ell))} - \alpha \, |y| \quad (9)$$

**Preference Transfer Training** For each input $x_i^\ell$ in language $\ell$ with its corresponding set of $N$ generated responses $\{y_{i1}^\ell, y_{i2}^\ell, \ldots, y_{iN}^\ell\}$, we assign scores using $\mathcal{R}_c$. The responses receiving the highest and lowest scores are then selected to construct a preference tuple $(x_i^\ell, y_{i+}^\ell, y_{i-}^\ell)$.

The multilingual preference dataset, denoted as $\mathcal{D}$, is constructed by aggregating all preference tuples across all languages:

$$\mathcal{D} = \bigcup_{\ell, i} (x_i^\ell, y_{i+}^\ell, y_{i-}^\ell) \quad (10)$$

Finally, we employ a negative log-likelihood (NLL) loss term for the chosen labels in DPO loss in Eq. (5) to improve multilingual alignment performance. The resulting optimization objective is formulated as:

$$\mathcal{L}_{\text{DPO}}^{\text{NLL}}(\pi_\theta) = -\frac{\log \pi_\theta(y_i^\ell | x_i^\ell)}{|y_i^\ell|}$$
$$- \log \sigma \left( \beta \log \frac{\pi_\theta(y_i^\ell | x_i^\ell)}{\pi_{\text{ref}}(y_i^\ell | x_i^\ell)} - \beta \log \frac{\pi_\theta(y_i^\ell | x_i^\ell)}{\pi_{\text{ref}}(y_i^\ell | x_i^\ell)} \right) \quad (11)$$

After DPO training, the policy model $\pi_\theta^t$ is updated to $\pi_\theta$, which is then used to generate responses and score data for the subsequent iteration. The overall process of our approach is illustrated in Algorithm 1 in Appendix A.2.

**Extension to Other DAA** KTO (Ethayarajh et al., 2024), inspired by prospect theory, directly optimizes generation utility, in contrast to DPO, which relies on pairwise preferences. We use an English KTO-aligned model as our base and apply KTO iteratively to explore the generalizability of our method beyond pairwise alignment. Details of the KTO optimization process with our approach can be found in Appendix A.6.

## 4 Discussion

In this section, we explore two key questions:
**1.** Are there alternative forms of implicit reward?
**2.** Is cross-lingual reward effective?

### 4.1 The Alternative Implicit Rewards

We designed alternative implicit rewards using a DPO-tuned model under the same settings and compared the effect of different rewards in Section 5.2 and Appendix C.4.

Prior work (Wu et al., 2024; Hong et al., 2024) shows that reward models trained only on English data can achieve *zero-shot cross-lingual transfer*. Therefore, the most straightforward reward approach is the multilingual reward. Given prompt $x_i^\ell$ and corresponding response $y$, the **multilingual reward** $\mathcal{R}_m$ is then calculated as:

$$\mathcal{R}_m = \beta \log \frac{\pi_\theta(y \mid x_i^\ell)}{\pi_{ref}(y \mid x_i^\ell)} - \alpha \, |y| \quad (12)$$

The alternative reward function directly leverages the English reward model by translating responses into English before applying the reward. Given prompt $x_i^\ell$ and corresponding response $y$, the **Translate-to-English reward** $\mathcal{R}_t$ is then calculated as:

$$\mathcal{R}_t = \beta \log \frac{\pi_\theta(\mathcal{T}(\ell, y) \mid x_i^{en})}{\pi_{ref}(\mathcal{T}(\ell, y) \mid x_i^{en})} - \alpha \, |y| \quad (13)$$

where the mapping function $\mathcal{T}(\ell, y)$ is defined as:

$$\mathcal{T}(\ell, y) = \begin{cases} y & \text{if } \ell = \text{en}, \\ \text{LLM-Translate}(y) & \text{if } \ell \neq \text{en}. \end{cases} \quad (14)$$

Here, $\mathcal{T}(\ell, y)$ acts as an identity function when $\ell$ is English, returning $y$. Otherwise, $\mathcal{T}(\ell, y)$ translates the response $y$ into English using the LLM's translation capabilities, the prompt is shown in E.2. **Critically**, the $\mathcal{R}_c$ and $\mathcal{R}_t$ rewards are *always* conditioned on the English instructions, either $\mathcal{G}(x_i^l)$ or $x_i^{en}$. This ensures that reward scoring across all languages is based on English instructions, keeping them within the reward model's effective range.

## 4.2 The Effectiveness of Cross-lingual Reward

To assess the effectiveness of the cross-lingual reward, we sampled 100 pairs per language from the preference pairs constructed by $\mathcal{R}_c$ and evaluated them using head-to-head comparisons with GPT-4o, the prompt is shown in E.3. Table 1 shows the resulting reward accuracy, demonstrating a strong positive signal across all languages.

Moreover, we assess the reward accuracy of multilingual reward $\mathcal{R}_m$, Translate-to-English reward $\mathcal{R}_t$ and natural multilingual reward model $\mathcal{R}_n$. We choose 'allenai/tulu-v2.5-13b-chatbot-arena-2023-rm' (Ivison et al., 2024) as $\mathcal{R}_n$, a reward model used for PPO training on the Chatbot Arena 2023 dataset (Chiang et al., 2024).

Table 1 demonstrates the zero-shot cross-lingual transfer capability of our English Implicit Reward Model, highlighting the effectiveness of Implicit Rewarding in English for multilingual instructions. We observed that evaluating translated non-English responses leads to unreliable reward scores. We hypothesize that this inaccuracy arises from assessing distorted translated data, consequently hindering performance in non-English languages. In contrast, our implicit cross-lingual reward approach exhibits a higher average reward accuracy across the evaluated languages compared to a reward model directly trained on natural multilingual preference data from Chatbot Arena.

| | Reward Accuracy (0-1) | | | | | |
|---|---|---|---|---|---|---|
| | en | es | ru | de | fr | **Avg** |
| $\mathcal{R}_c$ | 0.71 | 0.61 | 0.62 | 0.67 | 0.69 | 0.66 |
| $\mathcal{R}_m$ | 0.71 | 0.56 | 0.57 | 0.54 | 0.57 | 0.59 |
| $\mathcal{R}_t$ | 0.71 | 0.46 | 0.47 | 0.46 | 0.48 | 0.52 |
| $\mathcal{R}_n$ | 0.65 | 0.62 | 0.61 | 0.62 | 0.63 | 0.63 |

Table 1: The reward accuracy of preference pairs.

## 5 Experiments

### 5.1 Experimental Setup

**Models** While prior work (Meta, 2024; Yang et al., 2024a) offers numerous English DPO-tuned instruction-following models, their RLHF training details are often closed. To ensure transparency, we use Llama-3-8B-SFT-DPO (Meng et al., 2024) as our initial English-aligned model. This model, derived from Meta-Llama-3-8B via SFT on UltraChat-200k (Ding et al., 2023) and DPO on UltraFeedback (Cui et al., 2024), follows the Zephyr training pipeline (Tunstall et al., 2023) using open-source data.

**Languages** English serves as our core training language, enabling both cross-lingual preference transfer and iterative self-improvement. Our main experiments focus on Spanish (es), Russian (ru), German (de), and French (fr) to observe cross-lingual preference alignment. We also evaluate several low-resource languages, including Bengali (bn), Swahili (sw), and Thai (th), to assess performance in low-resource settings.

**Datasets** UltraFeedback (Cui et al., 2024) is a large-scale, high-quality AI feedback dataset comprising 60K preference samples closely aligned with human preferences. We randomly sampled 3K UltraFeedback's prompts and translated them into other languages using the Google Translate API to create parallel multilingual prompts.

**Implementation Details** We sample $N = 10$ responses per prompt using a *temperature* of $0.9$ and *top-p* of $1.0$ and optimized $\alpha$ to minimize the length difference between the chosen and rejected responses. See Appendix B.3 for further details.

**Evaluation and Metrics** We evaluated multilingual preference alignment from three aspects:
(1) First, we used **X-AlpacaEval Leaderboard** in Yang et al. (2024c), a multilingual extension of AlpacaEval 2.0 (Li et al., 2023b), to compare the multilingual instruction-following abilities of various models. To mitigate length bias in LLM preferences, we report both standard Win Rate (WR) and length-controlled (LC) Win Rates.
(2) Second, we used **Multilingual MT-Bench**, a multilingual adaptation of MT-Bench (Zheng et al., 2024a), which consists of open-ended questions designed to assess conversational and instruction-following skills. GPT-4o was used to score model responses on a scale of 1 to 10.

(3) Finally, to assess the alignment tax, we evaluated our model on **Multilingual NLP benchmarks**, including multilingual version of MMLU (Hendrycks et al., 2020), HellaSwag (Zellers et al., 2019), ARC Challenge (Clark et al., 2018), and TruthfulQA (Lin et al., 2021).

## 5.2 Main Results

**X-AlpacaEval Leaderboard** Table 2 shows that implicit cross-lingual rewarding enables continuous improvement in multilingual preference alignment across iterations. Average length-controlled (LC) and standard win rates (WR) increased by 5.97% and 12.72%, respectively. Furthermore, the English LC win rate steadily improves from 17.24% to 21.19%, confirming the effectiveness of implicit preference rewarding for bootstrapping English proficiency, as observed in (Kim et al., 2024; Chen et al., 2024a). This continuous improvement in English performance strengthens the implicit cross-lingual reward, which is crucial for our method's iterative optimization. Remarkably, our model, trained without any manually annotated multilingual preference data, outperforms similarly sized Instruct models, including Llama-3-8B-Instruct, Qwen2-7B-Instruct, and Mistral-7B-Instruct-v0.3 (LC: 18.24% vs. 13.74%, 18.10%, 17.29%), all of which were trained with extensive annotated preference data.

**Multilingual MT-Bench** The MT-Bench results in Table 15 show a continual performance improvement, increasing from 6.20 for $\pi_\theta^0$ to 6.77 for $\pi_\theta^2$. This improvement stems from the strong reward signal provided by implicit cross-lingual rewarding. Because we use GPT-4o as the reference model, its advanced capabilities result in lower *absolute* MT-Bench scores compared to GPT-4 evaluation. However, we focus on *relative* score changes during iterative training.

**Multilingual NLP Benchmarks** To assess the potential degradation of world knowledge and commonsense reasoning during alignment, known as the "alignment tax", Table 4 presents average results across the five training languages on four benchmarks (detailed results in Appendix C.1). The benchmark results show no performance degradation compared to the base model, indicating that our method effectively avoids introducing the alignment tax during preference optimization.

**Comparison** Yang et al. (2024c) proposed Language Imbalance Driven Rewarding, using language imbalance as a reward signal. We compare this approach to the same settings on X-AlpacaEval (Table 2). Note that we report the best model performance from two iterations, as we observed performance degradation in most languages in the second iteration of this approach. While it improves multilingual alignment over $\pi_\theta^0$, its gains are significantly smaller than ours. We attribute this to its reliance on language imbalance and self-translation, limiting its effectiveness. Moreover, it doesn't address length bias, resulting in limited *LC* gains. Further analysis is provided in the Appendix A.5.

**Extension to Other DAA** We extend our approach beyond DPO-aligned models to other Direct Alignment Algorithm (DAA), using an English KTO-aligned model as the base and applying KTO for iterative training. Results in Table 2 show our approach generalizes well to KTO-aligned models, effectively leveraging KTO for iterative optimization. A detailed analysis is provided in Appendix A.6.

**Different Implicit Rewards** To investigate the impact of implicit rewards on multilingual preference alignment, we compare the one-iteration performance of $\pi_\theta^1$ trained with three different implicit reward models on X-AlpacaEval (Figure 3).
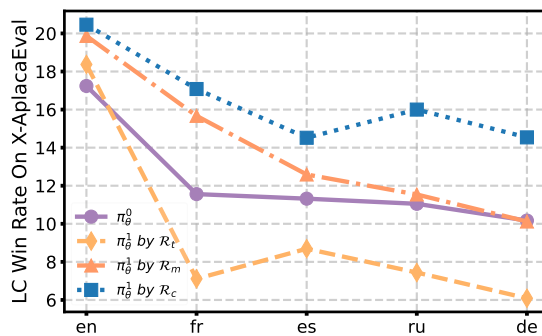


Figure 3: Improvement with different reward models.

The results reveal the following key findings: (1) The cross-lingual reward $\mathcal{R}_c$ yields the greatest improvement across all languages (3.20% to 5.52% in Table 12). (2) The multilingual reward $\mathcal{R}_m$ demonstrates effectiveness across most languages, suggesting zero-shot cross-lingual transfer of preference alignment. However, the effectiveness of this reward is highly dependent on the model's initial proficiency in a given language (as shown in Table 2). Consequently, as the model's initial proficiency decreases, the improvements con-

| Model | en LC | en WR | es LC | es WR | ru LC | ru WR | de LC | de WR | fr LC | fr WR | Avg LC | Avg WR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Cross-lingual Implicit Rewarding* | | | | | | | | | | | | |
| Llama-3-8B-SFT ($\pi_I$) | 9.02 | 6.25 | 6.34 | 3.77 | 3.96 | 3.28 | 3.71 | 2.62 | 4.73 | 3.26 | 5.55 | 3.84 |
| Llama-3-8B-SFT-DPO ($\pi_\theta^0$) | 17.24 | 17.35 | 11.32 | 12.41 | 11.05 | 13.82 | 10.17 | 11.87 | 11.56 | 13.09 | 12.27 | 13.71 |
| Iteration 1 ($\pi_\theta^1$) | 20.46 | 26.40 | 14.52 | 19.49 | 16.00 | 22.50 | 14.54 | 19.69 | 17.08 | 21.20 | 16.52 | 21.86 |
| Iteration 2 ($\pi_\theta^2$) | 21.19 | 31.38 | 16.88 | 23.37 | 18.11 | 25.76 | 17.92 | 26.27 | 17.12 | 25.35 | 18.24 | 26.43 |
| Meta-Llama-3-8B-Instruct | 23.48 | 24.90 | 17.52 | 18.08 | 6.37 | 7.81 | 7.74 | 8.65 | 13.58 | 14.18 | 13.74 | 14.72 |
| *Comparison: Language Imbalance Driven Rewarding* (Yang et al., 2024c) | | | | | | | | | | | | |
| Best Model of Two Iterations | 18.69 | 20.97 | 13.99 | 16.69 | 12.68 | 16.60 | 11.31 | 15.22 | 12.86 | 15.54 | 13.91 | 17.00 |
| *Extension to Other English DAA-aligned Model* | | | | | | | | | | | | |
| Llama-3-8B-SFT-KTO ($\pi_\theta^0$) | 14.99 | 15.86 | 13.21 | 14.22 | 10.72 | 14.74 | 10.14 | 12.18 | 11.55 | 13.49 | 12.12 | 14.10 |
| Iteration 1 ($\pi_\theta^1$) | 15.31 | 19.71 | 15.34 | 17.02 | 14.51 | 19.10 | 12.45 | 15.86 | 14.82 | 17.36 | 14.49 | 17.81 |
| Iteration 2 ($\pi_\theta^2$) | 15.19 | 21.36 | 15.39 | 16.60 | 16.13 | 19.47 | 14.47 | 17.26 | 15.25 | 17.22 | 15.29 | 18.38 |
| *SOTA Multilingual Models* | | | | | | | | | | | | |
| gpt-4o-mini | 47.33 | 45.17 | 48.56 | 44.63 | 48.53 | 47.03 | 48.54 | 44.20 | 48.03 | 44.93 | 48.20 | 45.19 |
| gpt-4-0613 | 28.86 | 15.61 | 35.08 | 18.18 | 30.37 | 16.82 | 29.10 | 16.00 | 25.44 | 15.23 | 29.77 | 16.37 |
| gpt-3.5-turbo-0125 | 24.50 | 11.96 | 31.79 | 14.42 | 28.21 | 13.74 | 27.82 | 12.41 | 28.71 | 12.70 | 28.21 | 13.05 |
| Qwen2-72B-Instruct | 39.56 | 37.72 | 36.43 | 24.73 | 37.38 | 27.15 | 32.51 | 23.93 | 33.47 | 24.63 | 35.87 | 27.63 |
| Meta-Llama-3-70B-Instruct | 36.54 | 39.74 | 30.65 | 32.58 | 7.43 | 9.14 | 8.26 | 9.48 | 23.27 | 25.20 | 21.23 | 23.23 |
| InternLM2.5-Chat-20B | 28.08 | 31.77 | 13.98 | 16.62 | 9.42 | 11.10 | 9.08 | 11.56 | 10.98 | 13.61 | 14.31 | 16.93 |
| Qwen2-7B-Instruct | 22.84 | 24.39 | 17.55 | 13.89 | 18.16 | 14.33 | 12.90 | 11.45 | 19.04 | 15.97 | 18.10 | 16.01 |
| Mistral-7B-Instruct-v0.3 | 25.13 | 21.46 | 16.30 | 13.36 | 14.16 | 13.75 | 14.48 | 11.91 | 16.37 | 13.28 | 17.29 | 14.75 |
| Aya-23-8B | 14.31 | 15.26 | 14.29 | 16.68 | 14.10 | 17.95 | 13.84 | 18.50 | 12.74 | 14.70 | 13.86 | 16.62 |

Table 2: **The X-AlpacaEval Leaderboard.** *LC* and *WR* denote length-controlled and standard win rate, respectively. The best and second-best scores in *Cross-lingual Implicit Rewarding* are highlighted in 'Green' and 'Lightgreen'. The X-AlpacaEval leaderboard was introduced by Yang et al. (2024c), which shows the win rate over GPT-4 Turbo evaluated by GPT-4.

| Model | Avg. Score (0-10) en | es | ru | de | fr | Avg |
|---|---|---|---|---|---|---|
| $\pi_\theta^0$ | 6.86 | 5.96 | 6.01 | 5.93 | 6.23 | 6.20 |
| $\pi_\theta^1$ | 6.93 | 6.61 | 6.42 | 6.76 | 6.56 | 6.66 |
| $\pi_\theta^2$ | 7.02 | 6.96 | 6.44 | 6.75 | 6.68 | 6.77 |

Table 3: The Multilingual MT-Bench Benchmark on Llama-3-8B-SFT-DPO, judged with GPT-4o.

ferred by the multilingual reward also diminish (as shown in Figure 3). This trend is evident in the performance gain, which ranges from 4.11% for French to near zero for German. **(3)** The translate-to-English reward, $\mathcal{R}_t$, degrades performance in all languages except English, suggesting that translating responses before reward evaluation is ineffective. We hypothesize that the translation process may distort the original meaning and context of the response, leading to inaccurate reward assignments and, consequently, reduced performance in non-English languages. **(4)** While English preference data is constant, English performance is still influenced by the preference data of other languages, emphasizing the importance of high-quality preference data for each language. Further analysis can

be found in Appendix C.4.

## 5.3 More Analysis

**Generalization to Lower-resource Languages** The strong performance on four high resource languages (*es, ru, de, fr*) naturally raises the question: *Can our method generalize to lower-resource languages?* Experiment with Bengali (bn), Swahili (sw), Thai (th), and English (en) in Table 5 shows the effectiveness of our approach in low-resource settings, demonstrating iterative performance gains across all languages. This is because implicit cross-lingual rewarding leverages the preference knowledge learned in English for direct (translation-free) reward, providing a strong, information-preserving reward signal for any language.

**Scaling the Number of Training Prompts** Figure 4 presents the X-AlpacaEval results for $\pi_\theta^0$ with varying training prompts in each language, demonstrating positive scaling with data volume. Notably, substantial improvements occur with as few as 1,000 prompts, a phenomenon aligned with the superficial alignment hypotheis (Zhou et al.,

| Model | Multilingual ARC challenge | Multilingual HellaSwag | Multilingual MMLU | Multilingual TruthfulQA | |
| | | | | MC1 | MC2 |
|---|---|---|---|---|---|
| Llama-3-8B-SFT | $0.4267_{\pm 0.0144}$ | $0.4974_{\pm 0.0051}$ | $0.5240_{\pm 0.0043}$ | $0.2909_{\pm 0.0161}$ | $0.4439_{\pm 0.0151}$ |
| Llama-3-8B-SFT-DPO ($\pi_\theta^0$) | $0.4653_{\pm 0.0145}$ | $0.5231_{\pm 0.0051}$ | $0.5349_{\pm 0.0043}$ | $0.3479_{\pm 0.0169}$ | $0.5079_{\pm 0.0161}$ |
| Iteration 1 ($\pi_\theta^1$) | $0.4679_{\pm 0.0145}$ | $0.5255_{\pm 0.0051}$ | $0.5356_{\pm 0.0043}$ | $0.3489_{\pm 0.0169}$ | $0.5099_{\pm 0.0162}$ |
| Iteration 2 ($\pi_\theta^2$) | $0.4674_{\pm 0.0145}$ | $0.5257_{\pm 0.0051}$ | $0.5364_{\pm 0.0043}$ | $0.3492_{\pm 0.0168}$ | $0.5092_{\pm 0.0163}$ |
| Meta-llama3-Instruct-8B | $0.4322_{\pm 0.0144}$ | $0.4833_{\pm 0.0051}$ | $0.5767_{\pm 0.0042}$ | $0.3403_{\pm 0.0168}$ | $0.5068_{\pm 0.0157}$ |

Table 4: The Multilingual NLP Benchmarks.

| Model | Win Rate | | | | Avg |
| | en | bn | sw | th | |
|---|---|---|---|---|---|
| $\pi_\theta^0$ | 17.35 | 4.35 | 3.43 | 14.17 | 9.83 |
| $\pi_\theta^1$ | 24.48 | 10.23 | 4.98 | 27.83 | 16.88 |
| $\pi_\theta^2$ | 32.06 | 14.09 | 6.28 | 29.55 | 20.50 |

Table 5: The X-AplacaEval Leaderboard On **Llama-3-8B-SFT-DPO** in **Lower-resource languages**.

2024). This highlights our method's efficiency and effective multilingual preference optimization with minimal data. Detailed results are provided in Appendix C.5.
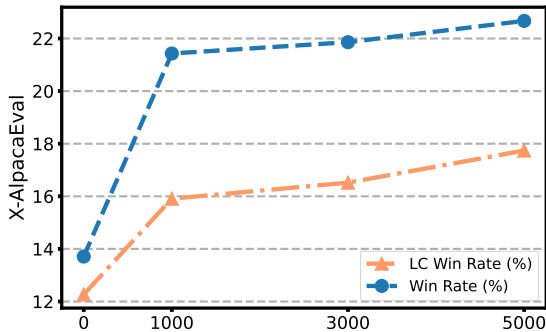


Figure 4: The average results for $\pi_\theta^0$ with varying training prompts.

**Diving into Implicit Cross-Lingual Reward** Table 6 investigates the impact of the cross-lingual reward defined in Eq. (9). **(1)** *Effect of Length Penalty:* By adding length penalty $\alpha|y|$ in reward, the generated responses became significantly shorter (2023.8 vs. 2474.4), while the LC win rate increased by 2.27%. While the standard win rate decreased, it is inherently susceptible to length bias. **(2)** *Effect of Reference Model Selection:* Compared to using the previous model $\pi_\theta^1$ as the reference, fixing the reference model to $\pi_I$ improves the length-controlled (LC) win rate from 16.74% to 18.24% while maintaining the win rate. Using $\pi_I$ as the reference ensures a less bias reward signal. When the previous model $\pi_\theta^1$ is used, the reward signal can become susceptible to the evolving preferences of the model itself. This can lead to the model ex-

ploiting spurious correlations, such as length bias, rather than focusing on genuine improvements in response quality. The less bias reward signal from $\pi_I$ mitigates this issue, allowing the model to focus on generating higher-quality responses, reflected in the increased *LC* win rate.

| Different Settings | LC | WR | Len |
|---|---|---|---|
| *Iteration 0: Initialization* | | | |
| Llama-3-8B-SFT ($\pi_I$) | 5.55 | 3.84 | 897.6 |
| Llama-3-8B-SFT-DPO ($\pi_\theta^0$) | 12.27 | 13.71 | 1695.2 |
| *Iteration 1: with / without Length Penalty $\alpha|y|$* | | | |
| ($\pi_\theta^1$, Eq. (9) without $\alpha|y|$) | 14.25 | 23.07 | 2474.4 |
| ($\pi_\theta^1$, Eq. (9) with $\alpha|y|$) | 16.52 | 21.86 | 2023.8 |
| *Iteration 2: with Different Reference Model* | | | |
| ($\pi_\theta^2$, Eq. (9) with $\pi_\theta^1$ as $\pi_{ref}$) | 16.74 | 26.62 | 2371.2 |
| ($\pi_\theta^2$, Eq. (9) with $\pi_I$ as $\pi_{ref}$) | 18.24 | 26.43 | 2254.6 |

Table 6: The Impact of Cross-lingual Reward.

## 6 Related Work

**Implicit Rewarding Optimization** Direct preference optimization (DPO) (Rafailov et al., 2024b) directly optimizes LLM to align with human preference by producing the optimal policy to an implicit reward model fit to the preference data. Rafailov et al. (2024a) proposed DPO within the token-level MDP setting, showing that it implicitly learns a token-level reward function using binary preference feedback. Zhong et al. (2024) introduced Reinforced Token Optimization (RTO) that performs PPO based on the implicit reward in DPO. Yang et al. (2024b); Chen et al. (2024b) use implicit reward margins predicted by DPO to efficiently annotate pairwise datasets. Chen et al. (2024a); Kim et al. (2024); Ko et al. (2024) utilized the implicit reward in the DPO-tuned model itself to construct a preference dataset and then used it in subsequent DPO rounds. Previous work has focused on using implicit rewards with English data in DPO-tuned models for English preference selection. Our work introduces implicit cross-lingual rewards, leveraging English DAAs-tuned models

to bootstrap capabilities across all languages.

**Multilingual Preference Alignment** Enhancing the multilingual capabilities of LLMs is crucial for enabling users worldwide to fully benefit from this technology (Yang et al., 2023; Li et al., 2023a; Liang et al., 2024). Prior work on multilingual rewarding (Wu et al., 2024; Hong et al., 2024) has explored cross-lingual transfer in reward model training using multilingual base models, showing zero-shot transfer capabilities. Due to multilingual preference data scarcity, Ahmadian et al. (2024); Dang et al. (2024) leveraged external, more powerful multilingual LLMs and reward models to construct multilingual preference data and applied optimization algorithms for multilingual alignment, incurring significant computational cost. MAPO (She et al., 2024) uses an external translation model as a reward model, aligning non-dominant languages with dominant ones by assessing consistency. However, the translator's limited context window may restrict it to other tasks. Zhao et al. (2025) introduced multilingual reward gap optimization to improve multilingual safety alignment. Yang et al. (2024c) utilized the inherent language imbalance within LLMs to generate rewards and self-improve multilingual performance; however, this approach yields relatively coarse reward signals. Our work addresses these limitations by using implicit cross-lingual rewarding with fine-grained reward signals to create paired data for self-iterative DPO training.

## 7 Conclusion

This paper proposes a simple yet effective framework that leverages the implicit reward model of English-aligned models as a fine-grained reward signal to bootstrap multilingual LLM alignment through a self-improving process. Our key insight is to directly leverage English-aligned models and introduce an implicit cross-lingual reward mechanism to generate preference labels, thereby explicitly capturing preference knowledge from aligned model. This labeled preference data is then used to fine-tune the model itself via direct alignment algorithms, enabling the transfer and refinement of preferences from English to other languages. Experimental results based on Llama3 demonstrate that our approach significantly improves multilingual preference alignment without any annotation data. This work offers a novel and efficient pathway for multilingual preference alignment.

## Limitations

Our work directly leverages the implicit cross-lingual reward derived from existing English-aligned models to iteratively improve the multilingual preference alignment of the model itself. The accuracy of the implicit cross-lingual reward significantly impacts the alignment effectiveness. If the reward signal is inaccurate or biased, it may lead to suboptimal preference optimization and hinder multilingual preference alignment. However, this is a common challenge in preference optimization, as RLHF also faces similar issues when the reward model is not accurate. Another limitation is that our work focuses on general multilingual preference alignment. Developing more language-specific alignment, such as cultural alignment, is an area we plan to explore in future work.

## Ethical Considerations

This work leverages the implicit reward model of English-aligned models as a fine-grained reward signal to bootstrap multilingual LLM alignment through a self-improving process, making a novel and significant contribution to multilingual preference alignment. This work is dedicated to the field of efficient multilingual preference alignment, improving the alignment of large models with human preferences in multiple languages, making them better used globally. Our contributions are entirely methodological. Therefore, this work does not have direct negative social impacts. In our experiments, we used publicly available datasets widely employed in prior research, containing no sensitive information to the best of our knowledge. The authors have followed ACL ethical guidelines, and the application of this work poses no apparent ethical risks.

## References

Arash Ahmadian, Beyza Ermis, Seraphina Goldfarb-Tarrant, Julia Kreutzer, Marzieh Fadaee, Sara Hooker, et al. 2024. The multilingual alignment prism: Aligning global and local preferences to reduce harm. *arXiv preprint arXiv:2406.18682*.

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

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.

Changyu Chen, Zichen Liu, Chao Du, Tianyu Pang, Qian Liu, Arunesh Sinha, Pradeep Varakantham, and Min Lin. 2024a. Bootstrapping language models with dpo implicit rewards. *arXiv preprint arXiv:2406.09760*.

Yifang Chen, Shuohang Wang, Ziyi Yang, Hiteshi Sharma, Nikos Karampatziakis, Donghan Yu, Kevin Jamieson, Simon Shaolei Du, and Yelong Shen. 2024b. Cost-effective proxy reward model construction with on-policy and active learning. *arXiv preprint arXiv:2407.02119*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, et al. 2024. Ultrafeedback: Boosting language models with scaled ai feedback. In *Forty-first International Conference on Machine Learning*.

John Dang, Arash Ahmadian, Kelly Marchisio, Julia Kreutzer, Ahmet Üstün, and Sara Hooker. 2024. Rlhf can speak many languages: Unlocking multilingual preference optimization for llms. *arXiv preprint arXiv:2407.02552*.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. 2024. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. 2022. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Jiwoo Hong, Noah Lee, Rodrigo Martínez-Castaño, César Rodríguez, and James Thorne. 2024. Cross-lingual transfer of reward models in multilingual alignment. *arXiv preprint arXiv:2410.18027*.

Hamish Ivison, Yizhong Wang, Jiacheng Liu, Zeqiu Wu, Valentina Pyatkin, Nathan Lambert, Noah A Smith, Yejin Choi, and Hanna Hajishirzi. 2024. Unpacking dpo and ppo: Disentangling best practices for learning from preference feedback. *Advances in neural information processing systems*, 37:36602–36633.

Dongyoung Kim, Kimin Lee, Jinwoo Shin, and Jaehyung Kim. 2024. Aligning large language models with self-generated preference data. *arXiv preprint arXiv:2406.04412*.

Jongwoo Ko, Saket Dingliwal, Bhavana Ganesh, Sailik Sengupta, Sravan Bodapati, and Aram Galstyan. 2024. Sera: Self-reviewing and alignment of large language models using implicit reward margins. *arXiv preprint arXiv:2410.09362*.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages

with reinforcement learning from human feedback. *arXiv preprint arXiv:2307.16039*.

Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2023a. Improving in-context learning of multilingual generative language models with cross-lingual alignment. *arXiv preprint arXiv:2311.08089*.

Chong Li, Wen Yang, Jiajun Zhang, Jinliang Lu, Shaonan Wang, and Chengqing Zong. 2024. X-instruction: Aligning language model in low-resource languages with self-curated cross-lingual instructions. *arXiv preprint arXiv:2405.19744*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023b. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.

Yupu Liang, Yaping Zhang, Cong Ma, Zhiyang Zhang, Yang Zhao, Lu Xiang, Chengqing Zong, and Yu Zhou. 2024. Document image machine translation with dynamic multi-pre-trained models assembling. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7084–7095.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*.

AI Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.

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

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024. Disentangling length from quality in direct preference optimization. *arXiv preprint arXiv:2403.19159*.

Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. 2024a. From $r$ to $q^*$: Your language model is secretly a q-function. *arXiv preprint arXiv:2404.12358*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024b. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Samyam Rajbhandari, Olatunji Ruwase, Jeff Rasley, Shaden Smith, and Yuxiong He. 2021. Zero-infinity: Breaking the gpu memory wall for extreme scale deep learning. In *Proceedings of the international conference for high performance computing, networking, storage and analysis*, pages 1–14.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shuaijie She, Shujian Huang, Wei Zou, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. Mapo: Advancing multilingual reasoning through multilingual alignment-as-preference optimization. *arXiv preprint arXiv:2401.06838*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Amos Tversky and Daniel Kahneman. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5:297–323.

Zhaofeng Wu, Ananth Balashankar, Yoon Kim, Jacob Eisenstein, and Ahmad Beirami. 2024. Reuse your rewards: Reward model transfer for zero-shot cross-lingual alignment. *arXiv preprint arXiv:2404.12318*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Sen Yang, Leyang Cui, Deng Cai, Xinting Huang, Shuming Shi, and Wai Lam. 2024b. Not all preference pairs are created equal: A recipe for annotation-efficient iterative preference learning. *arXiv preprint arXiv:2406.17312*.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtranslate: Augmenting large language models with multilingual translation capability over 100 languages. *arXiv preprint arXiv:2305.18098*.

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. 2024c. Language imbalance driven rewarding for multilingual self-improving. *arXiv preprint arXiv:2410.08964*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Zhihan Zhang, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. 2023. Plug: Leveraging pivot language in cross-lingual instruction tuning. *arXiv preprint arXiv:2311.08711*.

Weixiang Zhao, Yulin Hu, Yang Deng, Tongtong Wu, Wenxuan Zhang, Jiahe Guo, An Zhang, Yanyan Zhao, Bing Qin, Tat-Seng Chua, et al. 2025. Mpo: Multilingual safety alignment via reward gap optimization. *arXiv preprint arXiv:2505.16869*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024a. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024b. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

Han Zhong, Guhao Feng, Wei Xiong, Xinle Cheng, Li Zhao, Di He, Jiang Bian, and Liwei Wang. 2024. Dpo meets ppo: Reinforced token optimization for rlhf. *arXiv preprint arXiv:2404.18922*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

# Appendix

## A Implicit Cross-Lingual Rewarding

### A.1 Base Model Setup

Prior work (Meta, 2024; Yang et al., 2024a) has provided numerous instruction-following models fine-tuned with DAAs on English preference data. However, the RLHF procedures for most of these models are not publicly disclosed, making it unclear whether they were trained with preference data from other languages during the DPO stage. To thoroughly explore the effectiveness of our approach, we choose Llama-3-8B-SFT-DPO[1] provided by Meng et al. (2024) as our initial English-aligned model. Meta-Llama-3-8B is fine-tuned on UltraChat-200k (Ding et al., 2023), resulting in Llama-3-8B-SFT[2]. This model is then further optimized using Direct Preference Optimization (DPO) on UltraFeedback (Cui et al., 2024), yielding the final model, Llama-3-8B-SFT-DPO. The training pipeline of Llama-3-8B-SFT-DPO follows the recipe of Zephyr (Tunstall et al., 2023) and is trained on open-resource data, ensuring a high level of transparency.

Furthermore, Meng et al. (2024) provides models optimized with other Direct Alignment Algorithm (DAA) under the same data and training recipe. We choose Llama-3-8B-SFT-KTO[3] as the base policy model to extend our approach to other English DAA-aligned models.

### A.2 Algorithm Overview

Algorithm 1 outlines our proposed Implicit Cross-lingual Rewarding framework. The algorithm takes as input an initial model ($\pi_I$), an English-aligned model ($\pi_\theta^0$) trained with DPO using $\pi_I$, the number of iterations ($T$), and a set of parallel multilingual prompts ($\mathcal{X}$). The core idea is to iteratively refine the multilingual preference alignment of an existing English-aligned model by leveraging its inherent English preference alignment. In each iteration $t$, preference data ($\mathcal{D}_t$) is synthesized using the implicit cross-lingual reward $\mathcal{R}_c$, derived from the previous iteration's model ($\pi_\theta^{t-1}$), the initial model ($\pi_I$). This data generation process involves calculating cross-lingual rewards (as detailed in Eq. 7, 9, and 10). Then, policy and reference

---

[1] https://huggingface.co/princeton-nlp/Llama-3-Base-8B-SFT-DPO

[2] https://huggingface.co/princeton-nlp/Llama-3-Base-8B-SFT

[3] https://huggingface.co/princeton-nlp/Llama-3-Base-8B-SFT-KTO

models are initialized. For each mini-batch sampled from the preference data, a training loss based on a refined DPO loss incorporating negative log-likelihood (NLL) (Eq. 11) is calculated. The model parameters are then updated using gradient descent. After processing all mini-batches, the model for the next iteration ($\pi_\theta^t$) is initialized with the updated parameters. This process repeats for $T$ iterations, and the final multilingual aligned model ($\pi_\theta^T$) is returned.

### A.3 Optimize the Length control $\alpha$ in reward

In our reward function $\mathcal{R}_c$, we incorporate a length penalty term, $\alpha|y|$, to discourage the generation of overly long outputs. Subtracting this term incentivizes the model to produce concise and appropriately sized responses. The hyperparameter $\alpha$ controls the strength of this penalty; larger values of $\alpha$ impose stronger penalties for longer outputs. Following the approach in Chen et al. (2024a),ine extend it to the multilingual setting and optimize $\alpha$ for each language $\ell$ by minimizing the expected difference in length between preferred ($y_+^\ell$) and dispreferred ($y_-^\ell$) responses within our dataset $\mathcal{D}$:

$$\hat{\alpha}^\ell = \arg\min_a |\mathbb{E}_{(x^\ell, y_+^\ell, y_-^\ell) \sim \mathcal{D}}(|y_+^\ell| - |y_-^\ell|)| \quad (15)$$

This optimization aims to find the $\alpha$ that best balances response quality and length.

### A.4 The Format of Different rewards

We present the data format for the cross-lingual reward, multilingual reward, and Translate-to-English reward in Figure 5, providing a detailed breakdown of how each reward is structured and utilized within our approach to facilitate multilingual preference alignment.

### A.5 Comparison with *Language Imbalance Driven Rewarding*

Yang et al. (2024c) proposed *Language Imbalance Driven Rewarding* for multilingual self-improving, where the inherent language imbalance between dominant and non-dominant languages within LLMs is leveraged as a reward signal. Then, using LLM itself mutually translates the dominant and non-dominant language responses to construct multilingual preference data. While the premise of language imbalance driven rewarding is compelling, and its effectiveness was demonstrated with the Llama-3-8B-Instruct model, this approach relies on the model's internal language imbalance and translation capabilities.

**Algorithm 1** Implicit Cross-lingual Rewarding

---

**Input:** Initial model $\pi_I$, $\pi_\theta^0$ is English-aligned model using DPO with $\pi_I$, Iterations $T$, Parallel multilingual prompts $\mathcal{X}$

---

**for** $t = 1$ **to** $T$ **do**
    Sampling responses $\mathcal{Y}$ with $\pi_\theta^{t-1}$ and $\mathcal{X}$ (Eq. 7)
    Synthesizing preference data $\mathcal{D}_t$ by score $\mathcal{Y}$ with $\mathcal{R}_c$, derived from $\pi_\theta^{t-1}$, $\pi_I$ (Eq. 7, 9 and 10)
    Initialization of policy and reference models $\pi_\theta \leftarrow \pi_\theta^{t-1}$, $\pi_{ref} \leftarrow \pi_\theta^{t-1}$
    **for** mini-batch $B \sim \mathcal{D}_t$ **do**
        Calculate training loss $\mathcal{L}_{\text{DPO}}^{\text{NLL}}(\pi_\theta)$ with refined DPO loss incorporating NLL (Eq. 11)
        Update model parameter: $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{DPO}}^{\text{NLL}}(\pi_\theta)$
    **end for**
    Initializing next iteration model $\pi_\theta^t$ with the updated parameters $\theta$
**end for**
**return** $\pi_\theta^T$

---
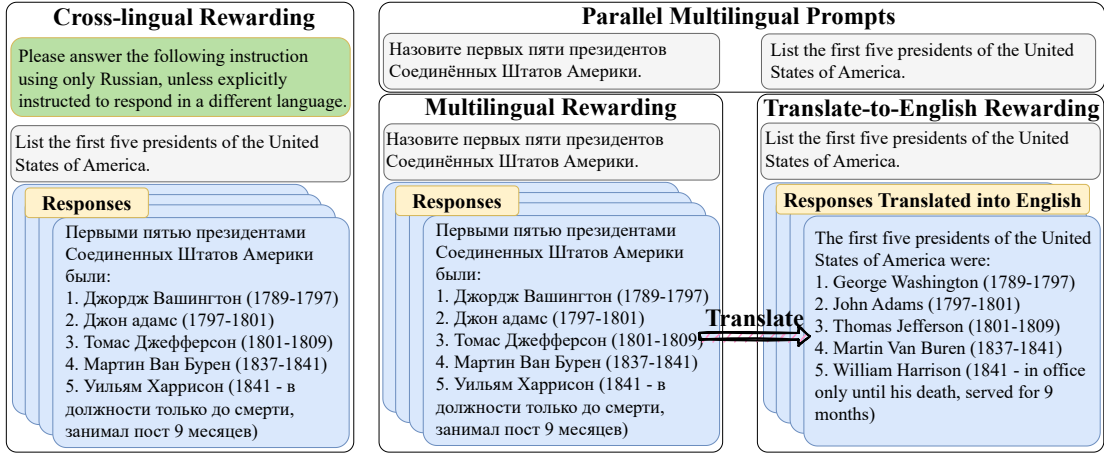


Figure 5: **The Format of Different Rewards.**

Table 7 highlights the differences in experimental settings between Language Imbalance Driven Rewarding (LIDR) and Implicit Cross-lingual Rewarding (ICR), including variations in the base model, training dataset, and dataset size. To ensure a more equitable and direct comparison of the methods, we conducted three distinct comparative experiments between our Implicit Cross-lingual Rewarding (ICR) approach and the Language Imbalance Driven Rewarding (LIDR) method.

First, a direct comparison of the reported results within their respective experimental settings (the upper of Table 8) shows that our method consistently outperforms LIDR excepted for English, achieving higher average performance across five training languages after only two iterations, even when using their original base models and configurations.

Then, we reproduced LIDR's results under our ICR experimental setup. These reproduced re-

sults (the middle of Table 8) reveal performance degradation for LIDR in most languages during the second iteration, resulting in substantially smaller overall gains compared to ICR.

We evaluated our approach under ICR's experimental conditions, specifically employing Llama3-8B-SFT as the base model within our ICR framework. As shown at the bottom of Table 8, our method consistently surpasses LIDR's performance even under their settings, demonstrating the robustness and effectiveness of the ICR approach.

## A.6 Extension to Other English DAA-aligned Model

Ethayarajh et al. (2024) proposed KTO, inspired by Kahneman and Tversky's prospect theory (Tversky and Kahneman, 1992), to directly maximize the utility of LLM generations rather than the log-likelihood of references. Unlike standard DPO and its variants, KTO eliminates the need for pair-

| Approach | Base Model | Training Dataset | Training Size |
|---|---|---|---|
| Language Imbalance Driven Rewarding | Llama3-8B-Instruct | Alpagasus | 1000 samples each language |
| Implicit Cross-lingual Rewarding | Llama3-8B-SFT-DPO | UltraFeedback | 3000 samples each language |

Table 7: Experimental settings between Language Imbalance Driven Rewarding (Yang et al., 2024c) and our approach.

| Model | en | | es | | ru | | de | | fr | | Avg | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | LC | WR | LC | WR | LC | WR | LC | WR | LC | WR | LC | WR |
| *Direct comparison under their respective settings.* | | | | | | | | | | | | |
| Llama3-8B-Instruct (Base Model) | 23.48 | 24.90 | 17.52 | 18.08 | 6.37 | 7.81 | 7.74 | 8.65 | 13.58 | 14.18 | 13.74 | 14.72 |
| LIDR (Iteration 1) | 26.10 | 30.11 | 18.78 | 21.82 | 14.23 | 18.01 | 14.36 | 16.87 | 14.49 | 17.51 | 17.59 | 20.86 |
| LIDR (Iteration 2) | 27.12 | 34.09 | 15.91 | 21.21 | 13.53 | 19.25 | 12.17 | 16.02 | 15.24 | 20.34 | 16.79 | 22.18 |
| Llama3-8B-SFT-DPO (Base Model) | 17.24 | 17.35 | 11.32 | 12.41 | 11.05 | 13.82 | 10.17 | 11.87 | 11.56 | 13.09 | 12.27 | 13.71 |
| ICR (Iteraion 1) | 20.46 | 26.40 | 14.52 | 19.49 | 16.00 | 22.50 | 14.54 | 19.69 | 17.08 | 21.20 | 16.52 | 21.86 |
| ICR (Iteraion 2) | 21.19 | 31.38 | 16.88 | 23.37 | 18.11 | 25.76 | 17.92 | 26.27 | 17.12 | 25.35 | 18.24 | 26.43 |
| *Comparison under LIDR settings.* | | | | | | | | | | | | |
| Llama3-8B-Instruct (Base Model) | 23.48 | 24.90 | 17.52 | 18.08 | 6.37 | 7.81 | 7.74 | 8.65 | 13.58 | 14.18 | 13.74 | 14.72 |
| LIDR (Iteraion 1) | 26.10 | 30.11 | 18.78 | 21.82 | 14.23 | 18.01 | 14.36 | 16.87 | 14.49 | 17.51 | 17.59 | 20.86 |
| LIDR (Iteraion 2) | 27.12 | 34.09 | 15.91 | 21.21 | 13.53 | 19.25 | 12.17 | 16.02 | 15.24 | 20.34 | 16.79 | 22.18 |
| ICR (Iteraion 1) | 26.80 | 31.76 | 21.92 | 24.03 | 16.03 | 19.31 | 16.79 | 19.37 | 18.95 | 21.49 | 20.10 | 23.19 |
| ICR (Iteraion 2) | 28.12 | 35.32 | 23.21 | 26.37 | 19.17 | 23.01 | 19.12 | 24.69 | 20.46 | 26.82 | 22.02 | 27.24 |
| *Comparison under ICR settings.* | | | | | | | | | | | | |
| Llama3-8B-SFT-DPO (Base Model) | 17.24 | 17.35 | 11.32 | 12.41 | 11.05 | 13.82 | 10.17 | 11.87 | 11.56 | 13.09 | 12.27 | 13.71 |
| LIDR* (Best of Two Iterations) | 18.69 | 20.97 | 13.99 | 16.69 | 12.68 | 16.60 | 11.31 | 15.22 | 12.86 | 15.54 | 13.91 | 17.00 |
| ICR (Iteraion 1) | 20.46 | 26.40 | 14.52 | 19.49 | 16.00 | 22.50 | 14.54 | 19.69 | 17.08 | 21.20 | 16.52 | 21.86 |
| ICR (Iteraion 2) | 21.19 | 31.38 | 16.88 | 23.37 | 18.11 | 25.76 | 17.92 | 26.27 | 17.12 | 25.35 | 18.24 | 26.43 |

Table 8: **Comparisons between Language Imbalance Driven Rewarding (LIDR) (Yang et al., 2024c) and Implicit Cross-lingual Rewarding (ICR) under various experimental settings on the X-AlpacaEval Leaderboard.** * notes that we report the best performance of the LIDR approach across two iterations under the ICR setting, as the second iteration showed performance degradation in most languages.

wise preferences, requiring only a binary signal indicating whether an output is desirable or undesirable for a given input. Therefore, we use an English KTO-aligned model as the base model and apply KTO for iterative optimization to investigate whether our method generalizes to non-pairwise direct alignment algorithms.

The KTO training loss is provided in the following:

$$
\begin{aligned}
\mathcal{L}(\pi_\theta) &= -\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\lambda_y - v(x,y)\right], \\
v(x,y) &= \begin{cases} \lambda_w \sigma(\beta \log \frac{\pi(y|x)}{\pi(y|x)} - z_{ref}), & \text{if } y \sim y_w|x, \\ \lambda_l \sigma(z_{ref} - \beta \log \frac{\pi(y|x)}{\pi(y|x)}), & \text{if } y \sim y_l|x. \end{cases} \\
z_{ref} &= \text{KL}(\pi_\theta(y|x)\|\pi_{\text{ref}}(y|x)).
\end{aligned}
\tag{16}
$$

where $\lambda_y$ denotes $\lambda_w$ for desirable response and $\lambda_l$ for undesirable response.

The implicit reward in KTO is derived in Eq. (16):

$$
r(x,y) = \log \frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)}
\tag{17}
$$

The reward function $r(x,y)$ derived in KTO is the same as that derived in DPO. Starting from Llama3-SFT-KTO, we use Algorithm 1, modifying the loss to the KTO loss, to perform iterative multilingual preference optimization based on the English KTO-aligned model.

The results in Table 2 show that after two iterations, the model's average Win Rate (WR) improved by 4.28%, and the average Length Control (LC) win rate improved by 3.17%. These results demonstrate the good generalization of our approach to other DAA-tuned models in English.

While multilingual preference optimization performs better starting from an English DPO-aligned

base model, due to differences in optimization algorithm performance and initial policy model capabilities, the effectiveness observed with KTO demonstrates that our approach can also achieve gains with weaker-aligned models.

## B Implementation Details

### B.1 Evaluation Details

**X-AlpacaEval Leaderboard** (Zhang et al., 2023) introduced the X-AlpacaEval benchmark, translated into Chinese, Korean, Italian, and Spanish by professional translators. Yang et al. (2024c) extended this benchmark to include German and Russian, and introduced the X-AlpacaEval Leaderboard, thereby expanding the original English-only AlpacaEval 2.0 (Li et al., 2023b) into a multilingual framework. We use the same prompts and configurations from X-AlpacaEval, as described in Yang et al. (2024c), to evaluate the multilingual instruction-following capabilities of LLMs. To mitigate length bias in LLM preferences, we report both standard and length-controlled (LC) win rates. The LC win rate is calculated using a separate regression model that isolates the impact of response quality by discounting the influence of length.

**Multilingual MT-Bench** (Zheng et al., 2024a) includes 80 open-ended questions that evaluate a chatbot's multi-turn conversational and instruction-following ability with human preference. We utilize the Multilingual MT-Bench from Yang et al. (2024c), which collected multilingual MT-Bench datasets including German, French, Russian, and Spanish. Specifically, we use `GPT-4o-2024-08-06` as our judge model and to generate reference outputs due to its advanced multilingual capabilities, ensuring more accurate evaluations. Because we use GPT-4o as the reference model, its advanced capabilities result in lower *absolute* MT-Bench scores compared to evaluations using GPT-4. However, our focus remains on the *relative* score changes observed throughout the iterative training process.

**Multilingual NLP Benchmark** We used the `lm-evaluation-harness` framework (Gao et al., 2024) to evaluate changes in world knowledge, commonsense reasoning, and honesty during the multilingual preference alignment iterations. Specifically, we chose the MMLU (Hendrycks

et al., 2020)[4], HellaSwag (Zellers et al., 2019)[5], ARC Challenge (Clark et al., 2018)[6] and TruthfulQA (Lin et al., 2021)[7] benchmarks, using the multilingual versions provided by Okapi (Lai et al., 2023). These multilingual benchmarks were created by translating the original benchmarks using ChatGPT. We list the detailed information of the benchmarks as follows:

**MMLU (Massive Multitask Language Understanding)**: This benchmark (Hendrycks et al., 2020) comprises 57 tasks, ranging from elementary math to law and ethics, testing a model's world knowledge and problem-solving abilities across diverse domains.

**HellaSwag**: HellaSwag (Zellers et al., 2019) is a challenging commonsense NLI benchmark focused on sentence completion. It presents multiple-choice questions where the plausible continuations require human-level commonsense inference. It is designed to be difficult for models relying on superficial statistical cues.

**The AI2 Reasoning Challenge (ARC) dataset**: The ARC dataset (Clark et al., 2018) focuses on question answering that contains questions from science exams from grade 3 to grade 9. It comprises two challenge sets: the Challenge Set, which contains the more difficult questions that require reasoning, and the Easy Set, which contains simpler questions.

**TruthfulQA**: TruthfulQA (Lin et al., 2021) evaluates a model's ability to measure whether a language model is truthful in generating answers to questions. It assesses whether a model can respond truthfully even when presented with misleading or deceptive information. Because evaluating truthfulness in generation tasks is difficult, the benchmark provides two multiple-choice formats, MC1 (single-true) and MC2 (multi-true), which test the model's ability to identify true statements.

### B.2 Experimental Environments

All experiments were conducted on 8 NVIDIA A800 80G GPUs. Our code primarily relies on Python 3.10 and PyTorch 2.3.0. Models were fine-tuned with LLaMA-Factory (Zheng et al.,

---

[4]`https://huggingface.co/datasets/alexandrainst/m_mmlu`
[5]`https://huggingface.co/datasets/alexandrainst/m_hellaswag`
[6]`https://huggingface.co/datasets/alexandrainst/m_arc`
[7]`https://huggingface.co/datasets/alexandrainst/m_truthfulqa`

2024b) and inference was performed with vLLM 0.6.1 (Kwon et al., 2023). Training for all models was launched with the accelerate (Gugger et al., 2022) library, utilizing DeepSpeed ZeRO Stage 2 (Rajbhandari et al., 2021).

## B.3 Hyperparameters

For preference pair construction, we sample $N = 10$ responses per prompt using a *temperature* of $0.9$ and *top-p* of $1.0$. During reward scoring, we optimized $\alpha$ to minimize the length difference between the chosen and rejected labels. For preference training, models are trained for one epoch per iteration with a learning rate of $5e-7$ and a batch size of $16$. The DPO hyperparameter $\beta$ was set to a fixed value of $0.1$ for all training runs. We employed the AdamW optimizer and a cosine learning rate scheduler with a warm-up phase corresponding to $3\%$ of the total training steps.

| Experiments | LR | BS | warm-up | Epoch | $\beta$ |
|---|---|---|---|---|---|
| *Cross-Lingual Rewarding* | | | | | |
| DPO-Tuned | 5e-7 | 16 | 0.03 | 1 | 0.1 |
| KTO-Tuned | 5e-7 | 32 | 0.03 | 1 | 0.1 |

Table 9: The hyperparameters on various experiments. 'LR' refers to the Learning Rate, and 'BS' denotes the Batch Size

## C Detailed Results and Analysis Across Languages

In this section, we provide more fine-grained results and analyses from our experiments to facilitate a clearer observation of each language's performance.

## C.1 Multilingual NLP Benchmark

Table 14 presents detailed results on four multilingual NLP benchmarks. These detailed results offer insights into our method's performance across various languages and tasks. The table demonstrates that our approach maintains performance comparable to the Llama-3-8B-SFT-DPO base model, effectively avoiding the "alignment tax" — the phenomenon where aligning a model with human preferences can negatively impact its performance on multilingual NLP tasks. This indicates that our approach successfully balances preference alignment with the preservation of general language understanding capabilities.

## C.2 Generalization to Lower-resource Languages

Table 10 presents performance results for lower-resource languages, including Bengali (bn), Swahili (sw), and Thai (th), which generally exhibit lower performance compared to middle-resource languages like Spanish, Russian, German, and French in Llama 3. While English saw a decline in length control win rate during the second iteration, possibly due to transferred length control preferences from other languages not perfectly aligned with optimal English preferences, the consistent win rate improvements across the other languages demonstrate the effectiveness of our cross-lingual implicit rewarding approach. This suggests that our method successfully transfers learned knowledge and preferences, promoting strong generalization even in lower-resource settings.

## C.3 Diving into Implicit Cross-Lingual Reward

Table 11 analyzes the effects of *Length Penalty* and *Reference Model Selection*.

To investigate the effect of *Length Penalty*, we compare controlled experiments in Iteration 1. Using optimal length penalty $\alpha|y|$ in cross-lingual rewarding minimizes the length difference between chosen and rejected responses, thereby reducing length bias in the preference data as much as possible. Compared to the setting without a length penalty, applying a length penalty improves the length control win rate across all languages. However, the shorter response length in the penalty setting also results in a slight decrease in the win rate across all languages except French.

Regarding *reference model selection*, using the initial model, $\pi_I$, as a fixed reference, instead of the previous iteration's model, $\pi_\theta^1$, further reduces average generation length (from 2371.2 to 2254.6) and consequently improves the length control (LC) win rate across all languages. While French (fr) and German (de) saw improvements in win rate, the other three languages experienced a slight decrease. Using the initial model ($\pi_I$) as a reference provides a less bias reward signal. A moving reference (like $\pi_\theta^1$) can lead to the reward signal drifting towards the model's own evolving (and potentially flawed) preferences, encouraging undesirable traits like length bias. The stability of a fixed $\pi_I$ mitigates this, promoting higher-quality responses and improving the length-controlled win rate.

| Model | en | | bn | | sw | | th | | Avg | |
|---|---|---|---|---|---|---|---|---|---|---|
| | LC | WR | LC | WR | LC | WR | LC | WR | LC | WR |
| *Cross-lingual Implicit Rewarding* | | | | | | | | | | |
| Llama-3-8B-SFT-DPO ($\pi_\theta^0$) | 17.24 | 17.35 | 2.95 | 4.35 | 2.61 | 3.43 | 10.29 | 14.17 | 8.27 | 9.83 |
| Iteration 1 ($\pi_\theta^1$) | 21.96 | 24.48 | 5.85 | 10.23 | 2.82 | 4.98 | 21.52 | 27.83 | 13.04 | 16.88 |
| Iteration 2 ($\pi_\theta^2$) | 17.68 | 32.06 | 7.85 | 14.09 | 3.21 | 6.28 | 23.19 | 29.55 | 12.98 | 20.50 |

Table 10: The X-ApacaEval Leaderboard on lower resource languages. *LC* and *WR* denote length-controlled and standard win rate, respectively.

| Model | en | | es | | ru | | de | | fr | | Avg | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | LC | WR | LC | WR | LC | WR | LC | WR | LC | WR | LC | WR | Len |
| Llama-3-8B-SFT-DPO ($\pi_\theta^0$) | 17.24 | 17.35 | 11.32 | 12.41 | 11.05 | 13.82 | 10.17 | 11.87 | 11.56 | 13.09 | 12.27 | 13.71 | 1695.2 |
| Iteration 1 ($\pi_\theta^1$) | 20.46 | 26.40 | 14.52 | 19.49 | 16.00 | 22.50 | 14.54 | 19.69 | 17.08 | 21.20 | 16.52 | 21.86 | 2023.8 |
| *Iteration 1: with / without Length Penalty $\alpha|y|$* | | | | | | | | | | | | | |
| $\pi_\theta^1$, Eq. (9) without $\alpha|y|$ | 16.78 | 27.05 | 13.98 | 21.12 | 15.04 | 24.95 | 12.76 | 21.45 | 12.70 | 20.78 | 14.25 | 23.07 | 2474.4 |
| $\pi_\theta^1$, Eq. (9) with $\alpha|y|$ | 20.46 | 26.40 | 14.52 | 19.49 | 16.00 | 22.50 | 14.54 | 19.69 | 17.08 | 21.20 | 16.52 | 21.86 | 2023.8 |
| *Iteration 2 with Different Reference Model* | | | | | | | | | | | | | |
| $\pi_\theta^2$, Eq. (9) with $\pi_\theta^1$ as $\pi_{ref}$ | 18.19 | 32.79 | 16.64 | 25.03 | 16.76 | 26.21 | 15.20 | 24.50 | 16.92 | 24.58 | 16.74 | 26.62 | 2371.2 |
| $\pi_\theta^2$, Eq. (9) with $\pi_I$ as $\pi_{ref}$ | 21.19 | 31.38 | 16.88 | 23.37 | 18.11 | 25.76 | 17.92 | 26.27 | 17.12 | 25.35 | 18.24 | 26.43 | 2254.6 |

Table 11: The X-AlpacaEval Leaderboard on the Analysis of Cross-Lingual Reward. **Len denotes the average character length of responses.**

## C.4 Different Implicit Rewards

To investigate the impact of different reward modeling on multilingual preference alignment, we compare the performance of $\pi_\theta^1$ trained for one iteration using three different types of implicit reward models on X-AlpacaEval. Table 12 presents the performance of $\pi_\theta^1$ on X-AlpacaEval under three reward modeling approaches: cross-lingual reward ($\mathcal{R}_c$), multilingual rewarding ($\mathcal{R}_m$), and Translate-to-English reward ($\mathcal{R}_t$).

The results reveal the following key findings:

(1) The cross-lingual reward $\mathcal{R}_c$ yields the greatest improvement in preference alignment across all languages, outperforming the other reward models. Furthermore, by leveraging the initial model's English-language reward capabilities, $\mathcal{R}_c$ confers substantial gains to $\pi_\theta^1$ across all languages, ranging from 3.20% to 5.52% shown in Table 12.

(2) The multilingual reward $\mathcal{R}_m$ demonstrates effectiveness across most languages, suggesting that preference alignment learned in English can be effectively transferred to other languages in a zero-shot manner, consistent with the findings of (Wu et al., 2024; Hong et al., 2024). However, the effectiveness of the multilingual reward is highly dependent on the model's initial proficiency in a given language. As the model's initial proficiency de-

creases, the improvements conferred by the multilingual reward also diminish. As shown in Figure 3 and Table 12, the improvement in $\pi_\theta^1$ conferred by $\mathcal{R}_m$ decreases as the initial model $\pi_\theta^0$'s alignment capability diminishes across languages, from 4.11% for French to near zero for German.

(3) The Translate-to-English reward, $\mathcal{R}_t$, leads to a performance decline in all languages except English, suggesting that translating responses into English before reward evaluation is ineffective. We hypothesize that this is because the implicit reward, derived from generation probabilities, is computed on parallel English data after translation. This translation process may distort the original meaning and context of the response, leading to inaccurate reward assignments and, consequently, reduced performance in non-English languages.

(4) While the English preference data remains consistent regardless of the reward model, performance differences arise during multilingual preference optimization. Although bootstrapping English preferences with implicit rewards is effective, as shown in prior work, our findings reveal that English performance is still influenced by preference data from other languages. Specifically, $\mathcal{R}_c$ achieves the best results, highlighting the importance of preference data quality across all languages when training multilingual models.

| Model | en | | es | | ru | | de | | fr | | Avg | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | LC | WR | LC | WR | LC | WR | LC | WR | LC | WR | LC | WR | Len |
| Llama-3-8B-SFT-DPO ($\pi_\theta^0$) | 17.24 | 17.35 | 11.32 | 12.41 | 11.05 | 13.82 | 10.17 | 11.87 | 11.56 | 13.09 | 13.23 | 13.37 | 1695.2 |
| *Iteration 1 with Different Reward Modeling* | | | | | | | | | | | | | |
| Translate-to-English $\mathcal{R}_t$, Eq. (13) | 18.37 | 24.95 | 8.69 | 12.56 | 7.44 | 12.01 | 6.09 | 10.33 | 7.11 | 12.63 | 12.41 | 13.78 | 2209.6 |
| Multilingual $\mathcal{R}_m$, Eq. (12) | 19.87 | 26.86 | 12.59 | 17.79 | 11.54 | 18.86 | 10.12 | 17.33 | 15.67 | 20.58 | 17.09 | 18.21 | 2119.6 |
| Cross-lingual $\mathcal{R}_c$, Eq. (9) | 20.46 | 26.40 | 14.52 | 19.49 | 16.00 | 22.50 | 14.54 | 19.69 | 17.08 | 21.20 | 18.92 | 20.25 | 2023.8 |

Table 12: The X-AlpacaEval Leaderboard on different Implicit Rewards.

## C.5 Scaling the Number of Training Prompts

Table 13 shows the effect of training set size on multilingual preference alignment performance. We can observe two points: (1) Increasing the training set size generally improved performance across most languages, although French (fr) showed signs of over-optimization when the number of training prompts rose from 3000 to 5000. (2) As the number of samples increases, the gain from the improvement becomes smaller. Using only 1000 prompts can improve LC and WR by 3.64% and 7.72%, respectively, while from 1000 to 5000, it only improves LC and WR by 1.83% and 1.24%. Our approach demonstrates efficient multilingual preference alignment, achieving strong performance with fewer training samples.

## D Dataset License

All models and data in our work are open-sourced. We utilize prompts from the UltraFeedback (Cui et al., 2024) dataset for efficient multilingual alignment. We adhere to the corresponding guidelines within the data.

| Model | en | | es | | ru | | de | | fr | | Avg | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | LC | WR | LC | WR | LC | WR | LC | WR | LC | WR | LC | WR | Len |
| Llama-3-8B-SFT-DPO ($\pi_\theta^0$) | 17.24 | 17.35 | 11.32 | 12.41 | 11.05 | 13.82 | 10.17 | 11.87 | 11.56 | 13.09 | 13.23 | 13.37 | 1695.2 |
| *Iteration 1 with Different Training Prompts in Each language* | | | | | | | | | | | | | |
| Llama-3-8B-SFT-DPO ($\pi_\theta^0$) | 17.24 | 17.35 | 11.32 | 12.41 | 11.05 | 13.82 | 10.17 | 11.87 | 11.56 | 13.09 | 12.27 | 13.71 | 1695.2 |
| $\pi_\theta^1$ with 1000 prompts | 19.93 | 23.74 | 15.27 | 20.71 | 14.19 | 20.82 | 13.84 | 19.91 | 16.34 | 21.96 | 15.91 | 21.43 | 2085.2 |
| $\pi_\theta^1$ with 3000 prompts | 20.46 | 26.40 | 14.52 | 19.49 | 16.00 | 22.50 | 14.54 | 19.69 | 17.08 | 21.20 | 16.52 | 21.86 | 2023.8 |
| $\pi_\theta^1$ with 5000 prompts | 24.19 | 26.96 | 16.71 | 21.09 | 16.24 | 22.49 | 16.55 | 21.75 | 15.01 | 21.05 | 17.74 | 22.67 | 2099.8 |

Table 13: The X-AlpacaEval results on Scaling the Number of Training Prompts.

| Model | Training Languages | | | | | Avg |
|---|---|---|---|---|---|---|
| | en | es | ru | de | fr | |
| *Multilingual ARC challenge, 0-shot* | | | | | | |
| Llama-3-8B-SFT | $0.5282_{\pm0.0146}$ | $0.4239_{\pm0.0145}$ | $0.3661_{\pm0.0141}$ | $0.3772_{\pm0.0142}$ | $0.4380_{\pm0.0145}$ | $0.4267_{\pm0.0144}$ |
| Llama-3-8B-SFT-DPO ($\pi_\theta^0$) | $0.5819_{\pm0.0144}$ | $0.4598_{\pm0.0146}$ | $0.3995_{\pm0.0143}$ | $0.4140_{\pm0.0144}$ | $0.4713_{\pm0.0146}$ | $0.4653_{\pm0.0145}$ |
| Iteration 1 ($\pi_\theta^1$) | $0.5742_{\pm0.0144}$ | $0.4684_{\pm0.0146}$ | $0.4021_{\pm0.0143}$ | $0.4234_{\pm0.0145}$ | $0.4713_{\pm0.0146}$ | $0.4679_{\pm0.0145}$ |
| Iteration 2 ($\pi_\theta^2$) | $0.5785_{\pm0.0144}$ | $0.4624_{\pm0.0146}$ | $0.4089_{\pm0.0144}$ | $0.4183_{\pm0.0144}$ | $0.4688_{\pm0.0146}$ | $0.4674_{\pm0.0145}$ |
| Meta-llama3-Instruct-8B | $0.5316_{\pm0.0146}$ | $0.4162_{\pm0.0144}$ | $0.3781_{\pm0.0142}$ | $0.3978_{\pm0.0143}$ | $0.4371_{\pm0.0145}$ | $0.4322_{\pm0.0144}$ |
| *Multilingual HellaSwag, 0-shot* | | | | | | |
| Llama-3-8B-SFT | $0.6008_{\pm0.0049}$ | $0.4997_{\pm0.0052}$ | $0.4412_{\pm0.0052}$ | $0.4600_{\pm0.0051}$ | $0.4855_{\pm0.0052}$ | $0.4974_{\pm0.0051}$ |
| Llama-3-8B-SFT-DPO ($\pi_\theta^0$) | $0.6292_{\pm0.0048}$ | $0.5270_{\pm0.0052}$ | $0.4624_{\pm0.0052}$ | $0.4864_{\pm0.0052}$ | $0.5104_{\pm0.0052}$ | $0.5231_{\pm0.0051}$ |
| Iteration 1 ($\pi_\theta^1$) | $0.6301_{\pm0.0048}$ | $0.5304_{\pm0.0052}$ | $0.4655_{\pm0.0052}$ | $0.4899_{\pm0.0052}$ | $0.5114_{\pm0.0052}$ | $0.5255_{\pm0.0051}$ |
| Iteration 2 ($\pi_\theta^2$) | $0.6295_{\pm0.0048}$ | $0.5306_{\pm0.0052}$ | $0.4655_{\pm0.0052}$ | $0.4922_{\pm0.0052}$ | $0.5105_{\pm0.0052}$ | $0.5257_{\pm0.0051}$ |
| Meta-llama3-Instruct-8B | $0.5764_{\pm0.0049}$ | $0.4877_{\pm0.0052}$ | $0.4326_{\pm0.0051}$ | $0.4483_{\pm0.0051}$ | $0.4715_{\pm0.0052}$ | $0.4833_{\pm0.0051}$ |
| *Multilingual MMLU, 5-shot* | | | | | | |
| Llama-3-8B-SFT | $0.6052_{\pm0.0039}$ | $0.5231_{\pm0.0043}$ | $0.4817_{\pm0.0044}$ | $0.4997_{\pm0.0043}$ | $0.5104_{\pm0.0044}$ | $0.5240_{\pm0.0043}$ |
| Llama-3-8B-SFT-DPO ($\pi_\theta^0$) | $0.6232_{\pm0.0039}$ | $0.5301_{\pm0.0043}$ | $0.4883_{\pm0.0044}$ | $0.5108_{\pm0.0043}$ | $0.5223_{\pm0.0044}$ | $0.5349_{\pm0.0043}$ |
| Iteration 1 ($\pi_\theta^1$) | $0.6236_{\pm0.0039}$ | $0.5293_{\pm0.0043}$ | $0.4853_{\pm0.0044}$ | $0.5103_{\pm0.0043}$ | $0.5297_{\pm0.0044}$ | $0.5356_{\pm0.0043}$ |
| Iteration 2 ($\pi_\theta^2$) | $0.6295_{\pm0.0039}$ | $0.5285_{\pm0.0043}$ | $0.4843_{\pm0.0044}$ | $0.5108_{\pm0.0043}$ | $0.5291_{\pm0.0044}$ | $0.5364_{\pm0.0043}$ |
| Meta-llama3-Instruct-8B | $0.6567_{\pm0.0038}$ | $0.5771_{\pm0.0043}$ | $0.5335_{\pm0.0044}$ | $0.5506_{\pm0.0043}$ | $0.5654_{\pm0.0043}$ | $0.5767_{\pm0.0042}$ |
| *Multilingual TruthfulQA MC1, 0-shot* | | | | | | |
| Llama-3-8B-SFT | $0.3060_{\pm0.0161}$ | $0.2725_{\pm0.0159}$ | $0.2919_{\pm0.0162}$ | $0.2779_{\pm0.0160}$ | $0.3062_{\pm0.0164}$ | $0.2909_{\pm0.0161}$ |
| Llama-3-8B-SFT-DPO ($\pi_\theta^0$) | $0.3856_{\pm0.0170}$ | $0.3232_{\pm0.0167}$ | $0.3452_{\pm0.0169}$ | $0.3363_{\pm0.0168}$ | $0.3494_{\pm0.0170}$ | $0.3479_{\pm0.0169}$ |
| Iteration 1 ($\pi_\theta^1$) | $0.3966_{\pm0.0171}$ | $0.3321_{\pm0.0168}$ | $0.3363_{\pm0.0168}$ | $0.3350_{\pm0.0168}$ | $0.3443_{\pm0.0169}$ | $0.3489_{\pm0.0169}$ |
| Iteration 2 ($\pi_\theta^2$) | $0.3896_{\pm0.0170}$ | $0.3370_{\pm0.0167}$ | $0.3378_{\pm0.0168}$ | $0.3385_{\pm0.0166}$ | $0.3433_{\pm0.0169}$ | $0.3492_{\pm0.0168}$ |
| Meta-llama3-Instruct-8B | $0.3611_{\pm0.0168}$ | $0.3333_{\pm0.0168}$ | $0.3541_{\pm0.0170}$ | $0.3173_{\pm0.0166}$ | $0.3355_{\pm0.0168}$ | $0.3403_{\pm0.0168}$ |
| *Multilingual TruthfulQA MC2, 0-shot* | | | | | | |
| Llama-3-8B-SFT | $0.4531_{\pm0.0147}$ | $0.4194_{\pm0.0150}$ | $0.4658_{\pm0.0157}$ | $0.4284_{\pm0.0150}$ | $0.4528_{\pm0.0152}$ | $0.4439_{\pm0.0151}$ |
| Llama-3-8B-SFT-DPO ($\pi_\theta^0$) | $0.5354_{\pm0.0158}$ | $0.4811_{\pm0.0162}$ | $0.5173_{\pm0.0164}$ | $0.4913_{\pm0.0160}$ | $0.5146_{\pm0.0162}$ | $0.5079_{\pm0.0161}$ |
| Iteration 1 ($\pi_\theta^1$) | $0.5460_{\pm0.0158}$ | $0.4848_{\pm0.0163}$ | $0.5163_{\pm0.0165}$ | $0.4931_{\pm0.0162}$ | $0.5094_{\pm0.0163}$ | $0.5099_{\pm0.0162}$ |
| Iteration 2 ($\pi_\theta^2$) | $0.5443_{\pm0.0159}$ | $0.4773_{\pm0.0164}$ | $0.5187_{\pm0.0166}$ | $0.4955_{\pm0.0163}$ | $0.5102_{\pm0.0164}$ | $0.5092_{\pm0.0163}$ |
| Meta-llama3-Instruct-8B | $0.5171_{\pm0.0152}$ | $0.4989_{\pm0.0157}$ | $0.5256_{\pm0.0162}$ | $0.4890_{\pm0.0157}$ | $0.5033_{\pm0.0158}$ | $0.5068_{\pm0.0157}$ |

Table 14: The Detailed Results of Multilingual NLP Benchmarks.

| Model | Avg. Score (0-10) | | | | | Avg |
|---|---|---|---|---|---|---|
| | en | es | ru | de | fr | |
| Llama-3-8B-SFT-DPO ($\pi_\theta^0$) | 6.86 | 5.96 | 6.01 | 5.93 | 6.23 | 6.20 |
| Iteration 1 ($\pi_\theta^1$) | 6.93 | 6.61 | 6.42 | 6.76 | 6.56 | 6.66 |
| Iteration 2 ($\pi_\theta^2$) | 7.02 | 6.96 | 6.44 | 6.75 | 6.68 | 6.77 |

Table 15: The Multilingual MT-Bench Benchmark on Llama-3-8B-SFT-DPO, judged with GPT-4o.

# E Prompt Template

## E.1 Cross-lingual Instruction Prefix P($\ell$) in mapping function $\mathcal{G}(x_i^\ell)$

> **Cross-lingual Instruction Prefix P($\ell$)**
>
> Please answer the following instruction using only $\ell$, unless explicitly instructed to respond in a different language.

## E.2 LLM-Translate($y$) in mapping function $\mathcal{T}(\ell, y)$

> **Prompt in LLM-Translate($y$)**
>
> Please translate the following sentences into *English*. The input sentences are wrapped by `<sentence>` and `</sentence>`:
>
> `<sentence>`
> y *(Response to $x_i^\ell$)*
> `</sentence>`

## E.3 Reward Accuracy Judgement Prompt

> **Prompt for Judging Reward Accuracy**
>
> You are a helpful following assistant whose goal is to select the preferred (least wrong) output for a given instruction in `[LANGUAGE]`.
>
> Which of the following answers is the best one for given instruction in `[LANGUAGE]`.
> A good answer should follow these rules:
> 1. It should be in `[LANGUAGE]`, except when the instruction explicitly requests the answer in a different language.
> 2. It should answer the request in the instruction.
> 3. It should be factually and semantically comprehensible.
> 4. It should be grammatically correct and fluent.
>
> `<instruction>`
> `[INSTRUCTION]`
> `</instruction>`
>
> `<answer1>`
> `[OUTPUT1]`
> `</answer1>`
>
> `<answer2>`
> `[OUTPUT2]`
> `</answer2>`
>
> FIRST, provide a one-sentence comparison of the two answers, explaining which you prefer and why.
> SECOND, on a new line, state only 'answer1' or 'answer2' to indicate your choice. If both answers are equally good or bad, state 'tie'. Your response should use the format:

Comparison: <one-sentence comparison and explanation>

Preferred: <'answer1' or 'answer2' or 'tie'>