MPO: Multilingual Safety Alignment via Reward Gap Optimization

Weixiang Zhao[1], **Yulin Hu**[1], **Yang Deng**[2], **Tongtong Wu**[3], **Wenxuan Zhang**[4]
Jiahe Guo[1], **An Zhang**[5], **Yanyan Zhao**[1*], **Bing Qin**[1], **Tat-Seng Chua**[5], **Ting Liu**[1]
[1]Harbin Institute of Technology, [2]Singapore Management University, [3]Monash University
[4]Singapore University of Technology and Design, [5]National University of Singapore
{wxzhao, yyzhao}@ir.hit.edu.cn

Abstract

Large language models (LLMs) have become increasingly central to AI applications worldwide, necessitating robust multilingual safety alignment to ensure secure deployment across diverse linguistic contexts. Existing preference learning methods for safety alignment, such as RLHF and DPO, are primarily monolingual and struggle with noisy multilingual data. To address these limitations, we introduce **M**ultilingual reward ga**P** **O**ptimization (**MPO**), a novel approach that leverages the well-aligned safety capabilities of the dominant language (*e.g.*, English) to improve safety alignment across multiple languages. MPO directly minimizes the reward gap difference between the dominant language and target languages, effectively transferring safety capabilities while preserving the original strengths of the dominant language. Extensive experiments on three LLMs, LLaMA-3.1, Gemma-2 and Qwen2.5, validate MPO's efficacy in multilingual safety alignment without degrading general multilingual utility. Our code is available at: https://github.com/circle-hit/MPO. WARNING: This paper may contain content that is offensive and harmful.

1 Introduction

Large language models (LLMs) are increasingly driving global applications (Brown et al., 2020; Touvron et al., 2023a,b; Jiang et al., 2023; Dubey et al., 2024; Team et al., 2024), enabling users from diverse linguistic and cultural backgrounds to access the benefits of AI advancements (Zhao et al., 2024b; Zheng et al., 2024a; Luo et al., 2024a; Xia and Luo, 2025). In this context, achieving multilingual safety alignment is crucial to ensuring secure deployment across various languages (Kanepajs et al., 2024; Friedrich et al., 2024). However, recent studies highlight substantial differences in the safety challenges faced by LLMs across various languages, with models being more prone to generate unsafe responses in low-resource languages. (Yong et al., 2023; Deng et al., 2024; Wang et al., 2024c; Shen et al., 2024).

To mitigate such challenge, one straightforward solution is to conduct safety preference alignment for each language, with methods like reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) or direct preference optimization (DPO) (Rafailov et al., 2023).

However, a key issue is the scarcity of multilingual data available (Ahmadian et al., 2024; Wu et al., 2024c; Hong et al., 2024a). Though off-the-shelf translation tools could be employed to generate training data in various languages, the resulting translations—especially for low-resource languages—are often noisy, riddled with unusual phrasing and inaccurate content (Zhang et al., 2024b; Liu et al., 2024a). On the other hand, current prevailing preference learning paradigms are highly sensitive to noisy data (Bai et al., 2022; Wang et al., 2024a; Chowdhury et al., 2024; Alfano et al., 2024). In some cases, such noise-induced errors may even cause safety misalignment (Shen et al., 2024; Razin et al., 2024), further exacerbating multilingual safety concerns.

To address this challenge, we first conduct an empirical analysis on several widely-used LLMs, including LLaMA-3.1 (Dubey et al., 2024), Gemma-2 (Team et al., 2024), and Qwen2.5 (Yang et al., 2024a), which have undergone sufficient safety alignment for their dominant language (typically English). We identify a crucial pattern: the implicit reward gap—defined as the log-likelihood difference between safe and unsafe responses—strongly correlates with multilingual safety performance. The dominant language (English) exhibits a substantially larger reward gap (RG) compared to low-resource ones, directly corresponding to its superior safety performance measured by Attack Success Rate (ASR). This inverse RG-ASR relationship es-

* Corresponding author

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 23564–23587

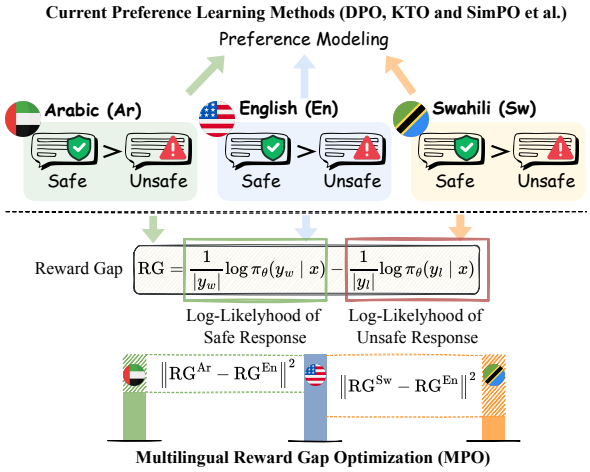


Figure 1: Top: Current preference learning methods optimize noisy multilingual preference data. Bottom: Our MPO directly minimizes the discrepancy of reward gap across different languages.

tablishes the reward gap as a quantifiable indicator of safety alignment quality across languages.

Building on these insights, we propose **M**ultilingual reward ga**P** **O**ptimization (**MPO**), a novel alignment paradigm for multilingual safety challenge that transfers safety capabilities from well-aligned dominant languages to others through reward gap optimization. As shown in Figure 1, unlike conventional preference learning approaches that attempt to directly optimize noisy multilingual preference data, MPO instead minimizes the discrepancy between the dominant language's robust and well-established reward gap and target languages' weaker alignment signals. To preserve the capabilities of the dominant language from degradation, we also incorporate constraints that maintain its hidden representations largely intact.

Our extensive experiments on LLaMA-3.1-8B-Instruct, Gemma-2-9B-it and Qwen2.5-7B-Instruct, showcase the efficacy and scalability of MPO in multilingual safety alignment over current preference learning methods without compromising the general multilingual utility. Deeper analysis reveals that MPO consistently outperforms across training datasets of varying quality. This further confirms that the reward gap of the dominant language serves as a more reliable and scalable supervision signal for effective multilingual safety alignment.

The main contributions of this work are summarized as follows:

- We propose to leverage the well-aligned safety capabilities of the dominant language as a high-quality supervision signal for multilingual safety alignment.
- We propose MPO, which directly minimizes the reward gap difference between the dominant language and target languages, enabling effective multilingual safety alignment.
- Experiments on three backbones demonstrate the superior performance of MPO over existing preference learning methods.

2 Preliminaries

In this section, we first introduce the formulation for the implicit reward gap re-parameterized by DPO (§2.1), as well as its improvements and optimizations in SimPO (§2.2). Specifically, we offer their corresponding interpretations in the context of ***multilingual safety alignment***.

2.1 Direct Preference Optimization (DPO)

DPO (Rafailov et al., 2023) is one of the most widely used methods for preference learning in LLM alignment. Unlike approaches that involve training an explicit reward model (Ouyang et al., 2022), DPO re-parameterizes the implicit reward function r using a closed-form expression derived from the Bradley-Terry (BT) model (Bradley and Terry, 1952) with the optimal policy:

$$r(x, y) = \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log Z(x), x \quad (1)$$

where π_θ is the policy model, π_{ref} is the reference model, typically the supervised fine-tuned (SFT) checkpoint, β is a hyper-parameter and $Z(x)$ is the partition function.

In the context of multilingual safety alignment, the reward gap of the backbone model between safe and unsafe responses in different languages t can be expressed as:

$$\begin{aligned} \text{RG}^t &= r(x^t, y_w^t) - r(x^t, y_l^t) \\ &= \beta \log \frac{\pi_\theta(y_w^t|x^t)}{\pi_{\text{ref}}(y_w^t|x^t)} - \beta \log \frac{\pi_\theta(y_l^t|x^t)}{\pi_{\text{ref}}(y_l^t|x^t)} \end{aligned} \quad (2)$$

where the triplet (x^t, y_w^t, y_l^t) are preference pairs related to safety concerns in language t, consisting of the input query x^t, the preferred (safe) response y_w^t, and the dispreferred (unsafe) response y_l^t.

2.2 Simple Preference Optimization (SimPO)

As pointed out by SimPO (Meng et al., 2024), using Eq. (1) as the implicit reward has the following drawbacks: it creates a mismatch between the reward optimized in training and the log-likelihood

		En	Zh	Ko	Ar	Bn	Sw
LLaMA-3.1	RG↑	**1.58**	0.36	0.29	0.60	0.04	0.05
	ASR↓	**9.00**	22.00	50.00	15.00	55.00	57.00
Gemma-2	RG↑	**2.32**	0.69	0.44	0.76	0.42	0.41
	ASR↓	**0.00**	9.00	14.00	4.00	24.00	26.00
Qwen-2.5	RG↑	**1.87**	**1.81**	0.69	0.78	0.14	0.20
	ASR↓	**13.00**	**9.00**	21.00	20.00	69.00	98.00

Table 1: Results of reward gap (RG) and safety performance across six languages. The evaluation metric used for safety is the Attack Success Rate (ASR), where lower values indicate better performance. Results of the dominant languages are highlighted in bold.

optimized during inference. To address this issue, SimPO considers using the average log-likelihood as the implicit reward:

$$r(x, y) = p_\theta(y \mid x) = \frac{1}{|y|} \log \pi_\theta(y \mid x) \tag{3}$$

Accordingly, the reward gap is formulated as:

$$\mathrm{RG}^t = \frac{1}{|y_w^t|} \log \pi_\theta(y_w^t \mid x^t) - \frac{1}{|y_l^t|} \log \pi_\theta(y_l^t \mid x^t) \tag{4}$$

We posit that, compared with Eq. (2), the reward gap in Eq. (4) provides a more accurate measure of safety performance differences across languages due to the following reasons: (1) It aligns with the likelihood metric that governs response generation, where a larger reward gap signifies a higher probability of producing safe responses over unsafe ones, serving as a direct indicator of safety performance. (2) Length normalization mitigates reward errors caused by length bias (Singhal et al., 2023; Park et al., 2024)—unsafe responses, which frequently include specific harmful content, are often longer than safe responses, which typically exhibit concise refusal patterns. Please refer to Appendix A for more emperical evidence and discussion.

3 Multilingual Reward Gap Optimization

In this section, we first demonstrate the relationship between the reward gap and the multilingual safety performance for different languages on three backbone LLMs (§3.1). Then, we derive the MPO objective (§3.2) and perform gradient analysis (§3.3).

3.1 Reward Gap across Languages

Models We select two English-centric LLMs: LLaMA-3.1-8B-Instruct (Dubey et al., 2024) and Gemma-2-9B-it (Team et al., 2024) and one bilingual LLMs: Qwen2.5-7B-Instruct (Yang et al., 2024a), to demonstrate the reward gap (Eq. (4)) on safety issues across different languages.

Languages We select six languages for evaluation based on the availability of language resources. The high-resource languages are English (En) and Chinese (Zh); the medium-resource languages are Korean (Ko) and Arabic (Ar); and the low-resource languages are Bengali (Bn) and Swahili (Sw). For LLaMA-3.1-8B-Instruct and Gemma-2-9B-it, En serves as the dominant language, while that for Qwen2.5-7B-Instruct is En and Zh.

Data We utilize the PKU-SafeRLHF dataset (Ji et al., 2024) for the reward gap evaluation across languages. This dataset comprises high-quality English preference pairs focused on safety-related questions. To extend its scope, we randomly sample 100 instances and translate them into each target language using the Google Translate API. Subsequently, we query LLMs directly with these multilingual inputs. The reward gap is computed using Eq. (4), while safety performance is evaluated based on the Attack Success Rate (ASR).

Analysis According to the results in Table 1, we can draw two key insights:

- ***Inverse relationship between RG and ASR***: Higher RG corresponds to lower ASR, indicating better safety performance. This demonstrates that RG can, to some extent, reflect the safety performance of LLMs in a specific language.
- ***Safety performance varies significantly across languages***: As reflected in RG values, lower-resource languages exhibit significantly lower RG compared to high-resource dominant ones, underscoring critical safety concerns in lower-resource settings.

3.2 The MPO Objective

Based on the above insights, we propose a novel method for multilingual safety alignment called **M**ultilingual reward ga**P** **O**ptimization (**MPO**). It takes the high-quality and well-aligned RG of the dominant language in LLMs as the pivot and aligns the RG of the target language to it. This facilitates the transfer of the dominant language's safety capabilities to the target language. This process can be formulated as:

$$\mathcal{L}_1 = \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\left\| \beta\, \mathrm{RG}^t - \mathrm{RG}^d \right\|^2 \right] \tag{5}$$

where t and d represent target and dominant languages, respectively. β functions to balance and stabilize the optimization (Haarnoja et al., 2018).

And RG is calculated by:

$$\mathrm{RG}^t = \frac{1}{|y_w^t|}\log \pi_\theta(y_w^t|x^t) - \frac{1}{|y_l^t|}\log \pi_\theta(y_l^t|x^t) \quad (6)$$

$$\mathrm{RG}^d = \frac{1}{|y_w^d|}\log \pi_{\text{ref}}(y_w^d|x^d) - \frac{1}{|y_l^d|}\log \pi_{\text{ref}}(y_l^d|x^d) \quad (7)$$

where the triplets (x^t, y_w^t, y_l^t) and (x^d, y_w^d, y_l^d) are preference pairs derived from target and dominant languages, respectively. Here, π_θ denotes the policy model, while π_{ref} serves as the reference model.

To ensure that the capabilities of the dominant language are not compromised, we constrain the representations of dominant language (at the position of the last token) to remain largely intact:

$$\mathcal{L}_2 = \mathbb{E}_{x^d \sim \mathcal{D}}\left[\left\|\boldsymbol{h}^d - \boldsymbol{h}^d_{\text{ref}}\right\|^2\right] \quad (8)$$

where $\boldsymbol{h}^d_{\text{ref}}$ is the representation of dominant language x^d obtained from the reference model. Inspired by recent empirical findings suggesting that modifying the hidden representations of LLMs is more effective for behavior control (Zou et al., 2023), we choose to constrain these representations directly, rather than applying KL-based regularization on logits (Ziegler et al., 2019).

The final optimization objective of MPO is:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 \quad (9)$$

3.3 What does the MPO update do?

The gradient for the learning of target languages with respect to the parameter θ can be written as:

$$\nabla_\theta \mathcal{L}_1(\theta) = 2\beta\, \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big(w_\theta \nabla_\theta \mathrm{RG}^t(\theta)\Big) \quad (10)$$

where $\nabla_\theta \mathrm{RG}^t(\theta)$ increases the likelihood of the preferred (safe) response y_w^t and decreases the likelihood of dispreferred (unsafe) response y_l^t for the target language, which is computed by:

$$\begin{aligned}\nabla_\theta \mathrm{RG}^t(\theta) = &\frac{1}{|y_w^t|}\nabla_\theta \log \pi_\theta(y_w^t \mid x^t) \\ &- \frac{1}{|y_l^t|}\nabla_\theta \log \pi_\theta(y_l^t \mid x^t)\end{aligned} \quad (11)$$

And we have $w_\theta = \beta\,\mathrm{RG}^t(\theta) - \mathrm{RG}^d$, which compares the reward gap between the target language $\beta\,\mathrm{RG}^t$ and the dominant language RG^d. This weight enables the model to adjust both the magnitude and direction of its gradient updates, while the extent of gradient descent is not dictated by the model's likelihood on the dataset. Thus, RG^d effectively sets the goal for how strongly the model should discriminate between y_w^t and y_l^t in target languages. Please refer to Appendix C for the derivation and detailed discussions.

4 Experiments

4.1 Experimental Setup

Models We use the same three backbones as in §3.1 to fully validate the efficacy and scalability of our MPO in safety alignment across languages.

Languages to be Safety Aligned We select six target languages, reflecting diverse linguistic families and resource levels. The high-resource languages are Chinese (Zh) and Japanese (Jp); the medium-resource languages are Korean (Ko) and Arabic (Ar); and the low-resource languages are Bengali (Bn) and Swahili (Sw). For LLaMA-3.1-8B-Instruct and Gemma-2-9B-it, English (En) serves as the dominant language, while that for Qwen2.5-7B-Instruct is En and Zh.

It is crucial to note that these target languages are deemed *out-of-scope* by the official model providers of the three backbones, who stress the importance of additional alignment efforts to guarantee safe and responsible deployment.

Training Data We sample 100 data points from PKU-SafeRLHF dataset (Ji et al., 2024) and translate them into each target language using the Google Translate API. This leads to that all methods are trained under the same 700 pairs of preference data. Details about the training data can be found in Appendix D. A comprehensive discussion on the effects of various translation tools and data volumes is provided in §5.2.

Benchmarks To comprehensively measure the efficacy of MPO on various safety scenarios, we employ 3 benchmarks for evaluation, including two multilingual jailbreak datasets: MultiJail (Deng et al., 2024) and Advbench-X (Yong et al., 2023), and one code-switch attack dataset: CSRT (Yoo et al., 2024). We use the Attack Success Rate (ASR) as our evaluation metric, calculated according to the evaluation pipeline proposed by Deng et al. (2024) with GPT-4o. Only meaningful refusal responses, excluding unrelated ones, are considered as failed attacks. As justified by Deng et al. (2024), this evaluation pipeline combining translation and GPT-4 classification achieves strong agreement with human annotations, with a Cohen's kappa score of 0.86, indicating a high level of alignment. Please refer to Appendix E for the detailed description of the evaluation setups.

Baseline Methods We compare MPO with supervised finetuning (**SFT**) (Ouyang et al., 2022)

	MultiJail							AdvBench-X								CSRT
	En	**Zh**	**Ko**	**Ar**	**Bn**	**Sw**	**AVG.**	**En**	**Zh**	**Jp**	**Ko**	**Ar**	**Bn**	**Sw**	**AVG.**	**-**
LLaMA-3.1	14.60	20.32	52.38	16.83	49.52	37.78	31.91	1.54	12.5	17.89	19.23	6.15	40.12	48.56	20.86	18.10
SFT	12.70	9.84	31.43	8.57	31.75	39.37	22.28	5.19	1.73	2.31	10.38	3.08	18.23	17.27	8.31	13.65
DPO	6.35	3.17	15.87	2.54	22.86	37.14	14.65	0.77	1.15	2.88	5.58	0.38	8.83	18.23	5.40	5.71
IPO	7.62	5.08	24.44	2.22	36.51	38.73	19.10	0.38	0.77	3.65	8.85	0.96	10.36	21.88	6.69	3.49
rDPO	15.24	14.13	44.29	18.73	50.79	56.83	33.34	6.35	5.77	3.85	11.54	8.08	60.65	56.62	21.84	11.43
CPO	22.85	41.26	29.21	38.10	66.98	66.98	44.23	1.35	2.69	3.85	5.78	1.35	20.96	29.23	9.32	19.37
KTO	4.76	6.67	21.59	4.76	30.79	42.86	18.57	0.58	0.96	3.27	8.46	1.92	11.35	22.84	7.05	7.31
ORPO	9.52	2.86	15.24	**1.27**	18.73	21.27	11.48	0.19	**0.00**	**0.19**	**1.35**	0.58	11.54	10.75	3.51	3.91
R-DPO	10.16	14.29	35.87	9.84	42.22	46.67	26.51	3.85	3.27	22.31	3.27	5.19	7.49	54.32	14.24	11.43
SimPO	9.21	8.25	30.48	7.30	40.63	42.22	23.02	5.77	3.46	11.73	17.69	5.19	28.94	21.25	13.43	7.62
MPO (Ours)	**2.22**	**0.95**	**4.76**	1.90	**12.38**	**10.79**	**5.98**	**0.00**	0.19	0.38	2.88	**0.00**	**7.10**	**5.37**	**2.27**	**1.59**
Gemma-2	2.54	9.52	14.61	4.13	20.32	14.60	10.95	0.96	1.15	3.08	5.00	3.85	6.72	5.18	3.71	4.76
SFT	2.86	4.44	13.02	4.76	23.17	12.38	10.11	**0.19**	**0.77**	1.92	4.42	2.50	5.00	4.22	2.72	5.74
DPO	2.23	7.30	10.79	6.35	23.82	13.33	10.64	0.38	1.73	1.54	3.46	3.08	5.03	3.84	2.72	5.71
IPO	2.86	8.89	16.19	5.08	18.41	14.92	11.06	0.77	1.54	2.50	4.42	3.65	8.25	5.18	3.76	6.37
rDPO	2.54	8.25	14.92	6.35	20.61	14.92	11.27	0.96	1.15	3.27	4.62	3.27	8.45	5.18	3.84	7.62
CPO	3.17	6.67	8.57	4.13	19.68	13.65	9.31	0.38	1.15	1.54	3.85	4.04	6.53	5.57	3.29	5.71
KTO	2.23	6.67	13.97	**3.49**	20.95	14.92	10.37	0.58	1.15	1.92	4.22	3.08	6.14	4.22	3.04	4.78
ORPO	3.17	6.03	10.16	5.71	17.14	10.48	8.78	0.38	1.54	0.96	2.88	**2.12**	5.84	4.26	2.57	6.67
R-DPO	3.81	7.62	12.70	6.35	28.25	13.97	12.12	0.58	1.92	4.42	4.81	3.46	7.68	4.80	3.95	6.03
SimPO	2.54	8.57	15.56	4.44	20.95	15.87	11.32	0.58	1.35	2.69	4.42	3.46	7.10	4.61	3.46	6.67
MPO (Ours)	**0.63**	**4.76**	**6.98**	3.81	**16.51**	**7.94**	**6.77**	0.38	0.96	**0.19**	**2.50**	2.69	**4.22**	**2.88**	**1.97**	**1.90**

Table 2: Detailed results on three multilingual safety benchmarks are presented. The evaluation metric used is the Attack Success Rate (ASR), where lower values indicate better performance. The best results achieved by our method and baselines are highlighted in bold, while the second-best results are underlined.

	MT-Bench		M-MMLU		MGSM	
	En	**Mul.**	**En**	**Mul.**	**En**	**Mul.**
LLaMA-3.1	7.31	4.81	67.70	45.35	88.00	40.13
+ MPO	7.25	4.92	67.10	44.67	88.00	44.67
Gemma-2	7.71	6.60	73.40	55.97	90.00	72.93
+ MPO	7.83	6.63	73.40	55.92	90.80	74.80

Table 3: Results of the multilingual utility evaluation. En denotes the performance of the dominant language, while Mul. represents the average performance across six target languages: Zh, Jp, Ar, Ko, Bn and Sw.

and the following preference optimization methods: **DPO** (Rafailov et al., 2023), **IPO** (Azar et al., 2024), **rDPO** (Chowdhury et al., 2024), **CPO** (Xu et al., 2024b), **KTO** (Ethayarajh et al., 2024), **ORPO** (Hong et al., 2024b), **R-DPO** (Park et al., 2024) and **SimPO** (Meng et al., 2024). Please refer to Appendix F for the detailed description of the baseline methods.

Implementation Details All training experiments are conducted on 8 A100 GPUs using the LLaMA-Factory (Zheng et al., 2024b). And our MPO is also implement based on this repo. For distributed training, we leverage the DeepSpeed (Rasley et al., 2020) with ZeRo-2 optimization. For more details, please refer to the Appendix G.

4.2 Overall Evaluation

Table 2 demonstrates the performance comparison of MPO and baselines based on LLaMA-3.1-8B-Instruct and Gemma-2-9B-it. Please refer to Appendix H.1 for more results on Qwen2.5-7B-Instruct. From the results across all backbones, we have drawn the following key insights:

MPO exhibits robust and consistent performance across various benchmarks and backbone models. It consistently surpasses all preference learning methods across three backbone LLMs and benchmarks, highlighting its outstanding safety alignment capabilities and scalability.

MPO excels in low-resource languages. Existing baseline methods often exhibit biased performance, disproportionately benefiting high-resource languages (e.g., Zh and Jp) and those where the model already demonstrates strong safety alignment (e.g., Ar). In contrast, MPO achieves comprehensive and significant improvements, particularly in low-resource languages (e.g., Bn and Sw). This highlights the effectiveness of leveraging high-quality internal safety alignment signals instead of relying exclusively on uneven preference data.

MPO maintains multilingual utility. Multilingual safety alignment should not compromise the

	MultiJail							MT-Bench							
	En	**Zh**	**Ko**	**Ar**	**Bn**	**Sw**	**AVG.**	**En**	**Zh**	**Jp**	**Ko**	**Ar**	**Bn**	**Sw**	**AVG.**
MPO	2.22	0.95	4.76	1.90	12.38	10.79	5.98	7.25	5.32	5.26	5.44	5.38	4.11	4.01	5.25
w/o Retain	2.23	0.63	1.90	0.95	10.16	13.33	**4.87**	7.19	5.09	4.41	5.27	4.79	3.46	3.79	4.86
w/ KL	14.60	22.54	58.73	22.54	58.10	75.87	42.06	7.41	5.33	5.16	5.58	5.38	4.28	4.11	5.32
w/o LN	17.78	26.67	57.46	26.67	65.08	77.14	45.13	7.41	5.61	5.29	5.53	5.38	4.31	4.31	**5.41**

Table 4: Ablation results on the key components of MPO. The best results are highlighted in bold.

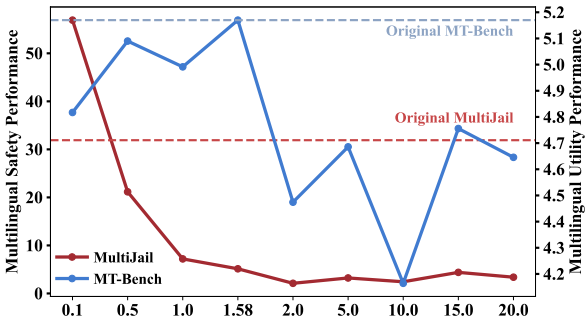


Figure 2: The results of replacing the dominant language reward gap with a fixed value on multilingual safety and general utility performance.

	MultiJail						
	En	**Zh**	**Ko**	**Ar**	**Bn**	**Sw**	**AVG.**
LLaMA-3.1	14.60	20.32	52.38	16.83	49.52	37.78	31.91
Align with Ar	6.98	6.67	20.00	4.13	17.78	46.35	16.99
Align with Bn	35.56	41.91	72.07	51.75	63.49	84.44	58.20
Align with Sw	20.63	30.16	53.97	26.35	53.97	81.90	44.50
MPO	**2.22**	**0.95**	**4.76**	**1.90**	**12.38**	**10.79**	**5.98**

Table 5: Multilingual safety performance when replacing reward gap with that from Ar, Bn and Sw as the supervision signal. The evaluation metric is the Attack Success Rate (ASR), where lower values indicate better performance. The best results are highlighted in bold.

model's multilingual general utility. Thus, we evaluate the resulting model across three key dimensions: (1) World Knowledge: M-MMLU (Hendrycks et al., 2021), (2) Reasoning: MGSM (Shi et al., 2023), and (3) Multi-turn Instruction-Following: MT-Bench (Zheng et al., 2023). The results in Table 3 show that MPO consistently maintains the general utility of both the dominant and target languages. For detailed results, evaluation settings and the comparison with baseline methods, please refer to Appendix H.2.

5 Analysis and Discussions

In this section, we offer a comprehensive analysis of MPO from: (1) ablation studies (§5.1), (2) the influence of preference data quality and quantity (§5.2), and (3) the rewards, representations, and case visualizations of the resulting model (§5.3). Unless stated otherwise, all analysis are conducted using the LLaMA-3.1 backbone.

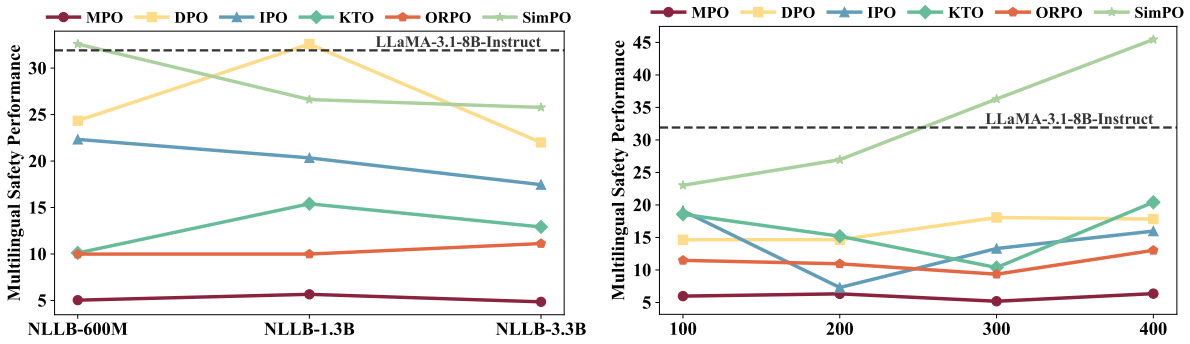
5.1 Ablation Study

Effect of Reward Gap from the Dominant Language as the Supervision Signal To assess the effectiveness of using the reward gap from the dominant language as an alignment objective, we conduct ablation experiments where we replace it with either a *fixed constant* or the *reward gap of other languages*. We also compare MPO against *recent cross-lingual transfer methods*.

Figure 2 shows the impact of replacing the dominant language reward gap with a fixed value (0.1–20) on multilingual safety and general utility. While increasing the constant enhances safety performance, it significantly degrades general utility due to excessive parameter shifts, leading to model collapse despite retention constraints. Notably, setting the constant to 1.58 (the training set's average reward gap of the dominant language) yields limited gains, highlighting the superiority of the fine-grained instance-level supervision in our MPO over coarse-grained dataset-level alignment. Please see Appendix H.3 for more details.

Table 5 further shows that using the reward gap of a target language as the alignment objective fails to yield meaningful safety improvements. Even when selecting the second-best safety-performing language (Ar) or low-resource languages (Sw, Bn), no effective multilingual safety enhancement is observed. This reinforces that the dominant language's reward gap provides a more reliable and high-quality supervision signal. For Qwen2.5, although it is a bilingual LLM with both Chinese and English as dominant languages, we find that using Chinese as the alignment target leads to better safety alignment performance compared to using English. Detailed results supporting this observation are provided in Appendix H.1, Table 9.

Table 14 in Appendix H.3 compares MPO



(a) Impact of Data Quality.

(b) Impact of Data Quantity.

Figure 3: Impact of the preference data. (a) Multilingual safety performance on MultiJail with varied data quality. (b) Multilingual safety performance on MultiJail with varied data size.

with state-of-the-art cross-lingual transfer methods, which align multilingual safety by either aligning multilingual representations: CLA (Li et al., 2024a) and LENS (Zhao et al., 2024a), or distilling knowledge from the dominant language: SDRRL (Zhang et al., 2024b). MPO consistently outperforms these methods, maintaining strong multilingual safety alignment. This further highlights the advantage of leveraging the dominant language's reward gap as a fine-grained supervision signal.

Effect of Other Components in MPO We further analyze the effect of other key components in Table 4. Removing the Retain component in Eq. (8) leads to a significant drop in multilingual utility, demonstrating its efficacy in preserving cross-lingual robustness. Introducing a KL-divergence-based constraint imposes a strong regularization that restricts the alignment of reward gap distributions across different languages, limiting the flexibility of MPO in adapting to multilingual safety preferences. Finally, removing length normalization (LN) in reward gap computation results in biased reward gap values, particularly in safety scenario that unsafe responses are often longer than safe ones, highlighting that LN effectively mitigates length-induced bias and facilitates more stable multilingual safety alignment. Please refer to Appendix H.3 for more detailed ablation studies, including our rationale for using the reference model—rather than the policy model—to compute the dominant language reward gap.

5.2 The Impact of Preference Data

Impact of Data Quality To evaluate the robustness of MPO across different levels of multilingual preference data quality, we employ three versions of the dataset obtained using three NLLB (Costa-jussà et al., 2022) translation models of varying sizes: NLLB-600M, NLLB-1.3B, and NLLB-3.3B. These models represent a progressive improvement in translation quality, with the largest model generally producing more accurate translations. Results are shown in Figure 3(a).

Baselines show considerable performance variations across different data quality levels, struggling to maintain stable safety alignment—even when trained on the highest-quality preference data (NLLB-3.3B). This underscores the challenges that noisy multilingual data pose for existing alignment methods. In contrast, MPO consistently delivers the best results across all data quality levels, demonstrating its stability and resilience to data noise. This validates the effectiveness of leveraging the reward gap in the dominant language as a source of high-quality supervision.

Further, recent studies explore LLMs themselves to generate multilingual preference data, rather than relying on external translation tools (She et al., 2024; Yang et al., 2024b, 2025). MPO consistently achieves the best multilingual safety alignment results data sources, demonstrating its robustness to variations in preference data. Please refer to Appendix H.4 for detailed results and analysis.

Impact of Data Quantity Figure 3(b) compares MPO with baseline methods across varying dataset sizes, with the x-axis representing the number of preference samples per language. MPO maintains stable performance across different data volumes, consistently outperforming baselines. However, all methods, including MPO, exhibit diminishing returns as data increases, with baseline performance even degrading with excessive data. This high-

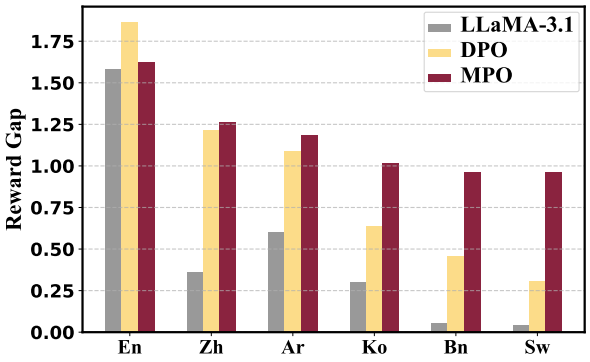


Figure 4: Reward gap across languages for the original backbone and those safety aligned by MPO and DPO.

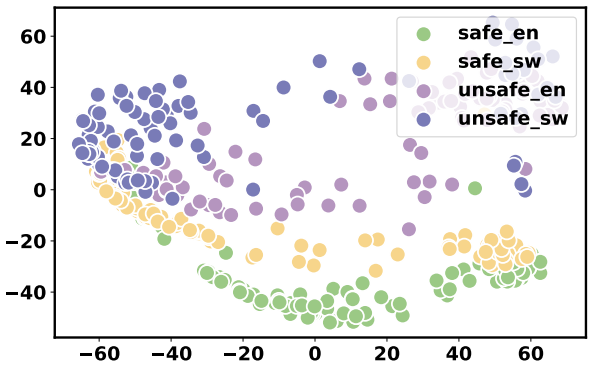


Figure 5: The visualization of multilingual representations for English and Swahili.

lights that enhancing supervision signal quality is far more effective than simply increasing data volume, aligning with broader LLM post-training trends (Zhou et al., 2023; Li et al., 2023; Cao et al., 2024; Guo et al., 2025; Ye et al., 2025).

5.3 Visualization Analysis

To better illustrate the impact of MPO on multilingual safety alignment, we visualize changes in the reward gap and the model's internal representation space. In Figure 4, MPO consistently achieves a higher reward gap than DPO across all languages. Notably, it significantly improves low-resource languages such as Swahili and Bengali, reducing the performance gap with English. Further, the visualization of the model's representation space in Figure 5, shows that MPO enables a clearer distinction between safe and unsafe responses in the target language Sw. This suggests that MPO enhances the model's ability to differentiate safety-critical responses, reinforcing its effectiveness in multilingual safety alignment. Please refer to Appendix H.5 for more visualization results.

6 Related Works

Multilingual Safety Vulnerability Recent studies have exposed risks in the multilingual safety of LLMs, underscoring the need for multilingual safety alignment (Qin et al., 2024; Li et al., 2024c; Gupta et al., 2024; Kanepajs et al., 2024; Verma and Bharadwaj, 2025). One line of approaches translate harmful prompts from high-resource to low-resource languages to assess safety (Yong et al., 2023; Deng et al., 2024; Xu et al., 2024b; Shen et al., 2024; Li et al., 2024b; Wang et al., 2024c; Poppi et al., 2024), as seen in Deng et al. (2024), which manually translated 315 English safety prompts (Ganguli et al., 2022) into nine languages. Others evaluate multilingual safety using code-switching, embedding multiple languages within the same harmful input (Gutiérrez-Clellen, 1999; Yoo et al., 2024; Song et al., 2024b; Upadhayay and Behzadan, 2024).

While these works have established a solid testbed for multilingual safety in LLMs, they have yet to introduce effective solutions to the existing challenges in this domain.

Safety Alignment Technique DPO (Rafailov et al., 2023) has emerged as a widely adopted offline preference learning method for aligning LLMs with human safety principles and values. In addition to DPO, various preference optimization objectives have been introduced. Ranking-based objectives enable comparisons among more than two instances (Dong et al., 2023; Yuan et al., 2023; Liu et al., 2024b; Song et al., 2024a). IPO (Azar et al., 2024) mitigates the overfitting issues inherent in DPO, while KTO (Ethayarajh et al., 2024) addresses preference optimization in non-pairwise data settings. Meanwhile, ORPO (Hong et al., 2024b) and SimPO (Meng et al., 2024) seek to remove reliance on a reference model.

Our proposed MPO stands out from existing methods in that we seek multilingual supervision signals from the *internal* reward gap of the LLMs, which specifically addresses the challenge of uneven data quality in multilingual safety alignment and offers fresh insights and new opportunities for achieving effective multilingual safety alignment.

7 Conclusion

In this paper, we introduce MPO, a novel approach to multilingual safety alignment that leverages the reward gap of the dominant language as a high-quality supervision signal. MPO directly mini-

mizes the discrepancy of reward gap across different languages to transfer safety alignment effectively. Experiments on LLaMA-3.1, Gemma-2, and Qwen2.5 confirm that MPO outperforms existing methods in multilingual safety alignment without compromising general multilingual utility. Further analysis shows that MPO remains robust across varying data qualities and sources, reinforcing the superiority of the dominant language's reward gap as a scalable alignment signal. These results establish MPO as a practical and effective solution for deploying multilingual-safe LLMs.

Limitations

This work has several limitations that provide directions for future research. Due to computational constraints, we conduct experiments on mid-scale models and did not extend our evaluation to larger-scale ones such as 32B even 72B LLMs. Future work should explore whether MPO scales effectively with larger models and whether its advantages persist at greater parameter sizes.

Additionally, we have focused exclusively on the application of MPO to multilingual safety alignment. However, there are more challenging and diverse alignment tasks that could be explored in the future, particularly those involving multicultural value alignment (Sorensen et al., 2024; Yao et al., 2024; Cahyawijaya et al., 2024). As multilingual safety alignment is only one aspect of broader ethical considerations, future work could extend the current methodology to tackle these value alignment challenges, ensuring models respect different cultural norms and ethical standards across regions.

Furthermore, given that safety guidelines are universal principles that users across various linguistic and cultural regions must adhere to, as emphasized in OpenAI (OpenAI, 2024b) and Meta's user guidelines (AI, 2024), it is reasonable to transfer the safety alignment of the dominant language to other languages. This idea has proven effective in our experiments, and we believe it could be validated in broader multilingual tasks in the future, particularly those that are language-agnostic, such as general problem-solving skills (Hu et al., 2024; Zhang et al., 2024a; Huang et al., 2024; Wang et al., 2024b; Luo et al., 2024b). Future work could explore these areas and broaden the scope of multilingual model evaluation, to ensure that advanced AI technologies are universally applicable and can promote responsible and ethical AI development on a global scale. We hope the research community continues to push forward in advancing these technologies and facilitating their global adoption.

Ethical Considerations

This work is conducted solely for academic research purposes and aims to address multilingual safety risks in large language models (LLMs). The primary goal of our study is to improve the robustness and consistency of LLMs across different languages, ensuring that they adhere to established safety principles regardless of linguistic variations. We acknowledge that multilingual safety alignment is a complex challenge, and our research does not aim to impose any specific cultural or ethical standards on diverse linguistic communities. Instead, our approach focuses on enhancing model consistency in following universally recognized safety guidelines, as outlined in user policies of major AI developers such as OpenAI and Meta. By ensuring equitable safety alignment across languages, we seek to mitigate risks associated with uneven safety performance in LLMs and reduce potential harm in lower-resource languages.

In conclusion, we aim to contribute to the development of fair, transparent, and globally applicable AI systems that align with responsible AI deployment principles. We encourage further community-driven research to refine multilingual safety alignment and promote the ethical and safe application of AI technologies worldwide.

Acknowledgments

We thank the anonymous reviewers for their comments and suggestions. This work was supported by the New Generation Artificial Intelligence-National Science and Technology Major Project 2023ZD0121100, the National Natural Science Foundation of China (NSFC) via grant 62441614 and 62176078, the Fundamental Research Funds for the Central Universities, and the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant (No. MSS24C012).

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. In *Proceedings of the 2024 Conference on Empir-*

ical Methods in Natural Language Processing, pages 12027–12049.

Meta AI. 2024. Meta safety policies. *Meta*.

Carlo Alfano, Silvia Sapora, Jakob Nicolaus Foerster, Patrick Rebeschini, and Yee Whye Teh. 2024. Learning loss landscapes in preference optimization. *arXiv preprint arXiv:2411.06568*.

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.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Samuel Cahyawijaya, Delong Chen, Yejin Bang, Leila Khalatbari, Bryan Wilie, Ziwei Ji, Etsuko Ishii, and Pascale Fung. 2024. High-dimension human value representation in large language models. *arXiv preprint arXiv:2404.07900*.

Boxi Cao, Keming Lu, Xinyu Lu, Jiawei Chen, Mengjie Ren, Hao Xiang, Peilin Liu, Yaojie Lu, Ben He, Xianpei Han, et al. 2024. Towards scalable automated alignment of llms: A survey. *arXiv preprint arXiv:2406.01252*.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024. Self-play fine-tuning converts weak language models to strong language models. In *Forty-first International Conference on Machine Learning*.

Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. 2024. Provably robust dpo: Aligning language models with noisy feedback. In *Forty-first International Conference on Machine Learning*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2024. Multilingual jailbreak challenges in large language models. In *The Twelfth International Conference on Learning Representations*.

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*.

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

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

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in english? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564.

Felix Friedrich, Simone Tedeschi, Patrick Schramowski, Manuel Brack, Roberto Navigli, Huu Nguyen, Bo Li, and Kristian Kersting. 2024. Llms lost in translation: M-alert uncovers cross-linguistic safety gaps. *arXiv preprint arXiv:2412.15035*.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, et al. 2024. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*.

Prannaya Gupta, Le Yau, Hao Low, I-Shiang Lee, Hugo Lim, Yu Teoh, Koh Hng, Dar Liew, Rishabh Bhardwaj, Rajat Bhardwaj, et al. 2024. Walledeval: A

comprehensive safety evaluation toolkit for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 397–407.

Vera F Gutiérrez-Clellen. 1999. Language choice in intervention with bilingual children. *American Journal of Speech-Language Pathology*, 8(4):291–302.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

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

Jiwoo Hong, Noah Lee, and James Thorne. 2024b. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189.

Peng Hu, Sizhe Liu, Changjiang Gao, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2024. Large language models are cross-lingual knowledge-free reasoners. *arXiv preprint arXiv:2406.16655*.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Yue Huang, Chenrui Fan, Yuan Li, Siyuan Wu, Tianyi Zhou, Xiangliang Zhang, and Lichao Sun. 2024. 1+ 1> 2: Can large language models serve as cross-lingual knowledge aggregators? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13394–13412.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Qiu, Boxun Li, and Yaodong Yang. 2024. Pku-saferlhf: Towards multi-level safety alignment for llms with human preference. *arXiv preprint arXiv:2406.15513*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Arturs Kanepajs, Vladimir Ivanov, and Richard Moulange. 2024. Towards safe multilingual frontier ai. In *Workshop on Socially Responsible Language Modelling Research*.

Sungdong Kim and Minjoon Seo. 2024. Rethinking the role of proxy rewards in language model alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20656–20674.

Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2024a. Improving in-context learning of multilingual generative language models with cross-lingual alignment. 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 8051–8069.

Jie Li, Yi Liu, Chongyang Liu, Ling Shi, Xiaoning Ren, Yaowen Zheng, Yang Liu, and Yinxing Xue. 2024b. A cross-language investigation into jailbreak attacks in large language models. *arXiv preprint arXiv:2401.16765*.

Yahan Li, Yi Wang, Yi Chang, and Yuan Wu. 2024c. Xtrust: On the multilingual trustworthiness of large language models. *arXiv preprint arXiv:2409.15762*.

Yunshui Li, Binyuan Hui, Xiaobo Xia, Jiaxi Yang, Min Yang, Lei Zhang, Shuzheng Si, Ling-Hao Chen, Junhao Liu, Tongliang Liu, et al. 2023. One-shot learning as instruction data prospector for large language models. *arXiv preprint arXiv:2312.10302*.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2024a. Is translation all you need? a study on solving multilingual tasks with large language models. *arXiv preprint arXiv:2403.10258*.

Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, et al. 2024b. Lipo: Listwise preference optimization through learning-to-rank. *arXiv preprint arXiv:2402.01878*.

Run Luo, Yunshui Li, Longze Chen, Wanwei He, Ting-En Lin, Ziqiang Liu, Lei Zhang, Zikai Song, Xiaobo Xia, Tongliang Liu, et al. 2024a. Deem: Diffusion models serve as the eyes of large language models for image perception. *arXiv preprint arXiv:2405.15232*.

Run Luo, Haonan Zhang, Longze Chen, Ting-En Lin, Xiong Liu, Yuchuan Wu, Min Yang, Minzheng Wang, Pengpeng Zeng, Lianli Gao, et al. 2024b. Mmevol: Empowering multimodal large language models with evol-instruct. *arXiv preprint arXiv:2409.05840*.

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

OpenAI. 2024a. Gpt-4o system card. *OpenAI*.

OpenAI. 2024b. Openai use policies. *OpenAI*.

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*.

Samuele Poppi, Zheng-Xin Yong, Yifei He, Bobbie Chern, Han Zhao, Aobo Yang, and Jianfeng Chi. 2024. Towards understanding the fragility of multilingual llms against fine-tuning attacks. *arXiv preprint arXiv:2410.18210*.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *arXiv preprint arXiv:2404.04925*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Leonardo Ranaldi, Giulia Pucci, Federico Ranaldi, Elena Sofia Ruzzetti, and Fabio Massimo Zanzotto. 2023. Empowering multi-step reasoning across languages via tree-of-thoughts. *arXiv preprint arXiv:2311.08097*.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 3505–3506.

Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, and Boris Hanin. 2024. Unintentional unalignment: Likelihood displacement in direct preference optimization. In *NeurIPS 2024 Workshop on Mathematics of Modern Machine Learning*.

Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacroce, Ahmed Awadallah, and Tengyang Xie. 2024. Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715*.

Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. MAPO: Advancing multilingual reasoning through multilingual-alignment-as-preference optimization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10015–10027.

Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024. The language barrier: Dissecting safety challenges of LLMs in multilingual contexts. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2668–2680.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2023. A long way to go: Investigating length correlations in rlhf. *arXiv preprint arXiv:2310.03716*.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2024a. Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18990–18998.

Jiayang Song, Yuheng Huang, Zhehua Zhou, and Lei Ma. 2024b. Multilingual blending: Llm safety alignment evaluation with language mixture. *arXiv preprint arXiv:2407.07342*.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Bibek Upadhayay and Vahid Behzadan. 2024. Sandwich attack: Multi-language mixture adaptive attack on llms. *arXiv preprint arXiv:2404.07242*.

Nikhil Verma and Manasa Bharadwaj. 2025. The hidden space of safety: Understanding preference-tuned llms in multilingual context. *arXiv preprint arXiv:2504.02708*.

Binghai Wang, Rui Zheng, Lu Chen, Yan Liu, Shihan Dou, Caishuang Huang, Wei Shen, Senjie Jin, Enyu Zhou, Chenyu Shi, et al. 2024a. Secrets of rlhf in large language models part ii: Reward modeling. *arXiv preprint arXiv:2401.06080*.

Weixuan Wang, Barry Haddow, Minghao Wu, Wei Peng, and Alexandra Birch. 2024b. Sharing matters: Analysing neurons across languages and tasks in llms. *arXiv preprint arXiv:2406.09265*.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael Lyu. 2024c. All languages matter: On the multilingual safety of LLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5865–5877.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Junkang Wu, Xue Wang, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. 2024a. α-dpo: Adaptive reward margin is what direct preference optimization needs. *arXiv preprint arXiv:2410.10148*.

Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. 2024b. Self-play preference optimization for language model alignment. *arXiv preprint arXiv:2405.00675*.

Zhaofeng Wu, Ananth Balashankar, Yoon Kim, Jacob Eisenstein, and Ahmad Beirami. 2024c. Reuse your rewards: Reward model transfer for zero-shot cross-lingual alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1332–1353.

Xiaobo Xia and Run Luo. 2025. Gui-r1: A generalist r1-style vision-language action model for gui agents. *arXiv preprint arXiv:2504.10458*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024a. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. In *Forty-first International Conference on Machine Learning*.

Nan Xu, Fei Wang, Ben Zhou, Bangzheng Li, Chaowei Xiao, and Muhao Chen. 2024b. Cognitive overload: Jailbreaking large language models with overloaded logical thinking. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3526–3548.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

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

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. 2025. Implicit cross-lingual rewarding for efficient multilingual preference alignment. *arXiv preprint arXiv:2503.04647*.

Jing Yao, Xiaoyuan Yi, Yifan Gong, Xiting Wang, and Xing Xie. 2024. Value fulcra: Mapping large language models to the multidimensional spectrum of basic human value. 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 8754–8777.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*.

Zheng Xin Yong, Cristina Menghini, and Stephen Bach. 2023. Low-resource languages jailbreak gpt-4. In *Socially Responsible Language Modelling Research*.

Haneul Yoo, Yongjin Yang, and Hwaran Lee. 2024. Code-switching red-teaming: Llm evaluation for safety and multilingual understanding. *arXiv preprint arXiv:2406.15481*.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. Rrhf: Rank responses to align language models with human feedback. *Advances in Neural Information Processing Systems*, 36.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. 2024. Self-rewarding language models. In *Forty-first International Conference on Machine Learning*.

Shimao Zhang, Changjiang Gao, Wenhao Zhu, Jiajun Chen, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2024a. Getting more from less: Large language models are good spontaneous multilingual learners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8037–8051.

Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. 2024b. Enhancing multilingual capabilities of large language models through self-distillation from resource-rich languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 11189–11204.

Weixiang Zhao, Yulin Hu, Jiahe Guo, Xingyu Sui, Tongtong Wu, Yang Deng, Yanyan Zhao, Bing Qin, Wanxiang Che, and Ting Liu. 2024a. Lens: Rethinking multilingual enhancement for large language models. *arXiv preprint arXiv:2410.04407*.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024b. Wildchat: 1m chatgpt interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, et al. 2024a. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024b. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

Zhenglin Zhou, Xiaobo Xia, Fan Ma, Hehe Fan, Yi Yang, and Tat-Seng Chua. 2025. Dreamdpo: Aligning text-to-3d generation with human preferences via direct preference optimization. *arXiv preprint arxiv:2502.04370*.

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

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

		En	Zh	Ko	Ar	Bn	Sw
LLaMA-3.1	**RG↑**	**27.02**	25.09	22.82	27.85	32.21	29.77
	ASR↓	**9.00**	22.00	50.00	15.00	55.00	57.00
Gemma-2	**RG↑**	**214.57**	63.92	45.54	88.46	80.62	96.60
	ASR↓	**0.00**	9.00	14.00	4.00	24.00	26.00
Qwen-2.5	**RG↑**	**13.84**	**12.31**	9.77	19.76	5.04	23.77
	ASR↓	**13.00**	**9.00**	21.00	20.00	69.00	98.00

Table 6: Results of reward gap (calculated by Eq. 2) and safety performance across six languages. The evaluation metric used for safety is the Attack Success Rate (ASR), where lower values indicate better performance. Results of the dominant languages are highlighted in bold.

A Further Discussion on Reward Gap

A.1 Definition

Our use of the term reward gap follows the theoretical foundations of Direct Preference Optimization (DPO) (Rafailov et al., 2023) and Simple Preference Optimization (SimPO) (Meng et al., 2024), where the log-likelihood is treated as a proxy for implicit reward. DPO derives this from the Bradley-Terry model (Bradley and Terry, 1952), and SimPO further simplifies it using average log-likelihood.

In our case, for each input, we compute the implicit reward for both the safe and unsafe responses using log-likelihoods. The difference between these two values is what we define as the reward gap (or margin). This gap reflects how much more the model prefers the safe response than unsafe one and is strongly correlated with multilingual safety performance.

A.2 Emperical and Theoritical Justification

We have already discussed in §2.2 that using Eq. (4) within SimPO (Meng et al., 2024) to compute the reward gap is more reasonable compared to Eq. (2) in DPO (Rafailov et al., 2023). Additionally, in §3.1, we provide an intuitive demonstration of the advantages of using Eq. (4) for reward gap calculation. Here, we conduct a more in-depth analysis to illustrate the limitations of Eq. (2).

Table 6 presents the results of computing the reward gap using Eq. (2) across three different backbone models. β is set to 1.0 and the base version of these backbones are adopted as the reference models. The dataset used for this evaluation remains consistent with that in §3.1. However, the results indicate that the computed reward gap fails to accurately reflect the model's safety performance across different languages. We attribute this discrepancy to the following three key reasons:

(1) **Inference-Training Objective Mismatch**: The reward formulation in Eq. (2) is derived from the implicit reward used during the training phase, but it does not directly align with the log-likelihood objective that governs inference (Meng et al., 2024). As a result, the reward gap computed with Eq. (2) may not faithfully capture the model's actual generation behavior, leading to misleading safety performance evaluations.

(2) **Bias in the Reference Model**: Ideally, the reference model used for computing the reward gap in a preference-optimized model should be the supervised fine-tuned (SFT) model from the previous training stage, rather than the base model (Ouyang et al., 2022; Rafailov et al., 2023). However, model providers do not publicly release this intermediate SFT model, making it difficult to obtain an accurate reference. As a result, using the base model as the reference introduces bias (Hong et al., 2024b; Wu et al., 2024a), further compromising the reliability of Eq. (2) in assessing safety performance.

(3) **Length Bias Effects**: Eq. (2) does not incorporate length normalization, making it susceptible to biases introduced by response length disparities (Meng et al., 2024; Kim and Seo, 2024). Empirically, unsafe responses tend to be longer due to the presence of explicit harmful content, while safe responses are often concise refusals. This discrepancy skews the reward gap calculations, causing inconsistencies in cross-linguistic safety evaluations.

These limitations collectively suggest that Eq. (2) from DPO (Rafailov et al., 2023) is not a reliable metric for evaluating safety differences across languages. In contrast, Eq. (4) from SimPO (Meng et al., 2024) mitigates these issues by normalizing the log-likelihood with sequence length, ensuring a more accurate measure of safety performance.

B Further Discussion on MPO

Here we conduct further explanation of how reward gap of the dominant language RG^d influences the learning of y_w^t and y_l^t for different languages.

Recall that:

$$\mathrm{RG}^d = \frac{1}{|y_w^d|} \log \pi_{\text{ref}}(y_w^d \mid x^d) - \frac{1}{|y_l^d|} \log \pi_{\text{ref}}(y_l^d \mid x^d),$$

where π_{ref} is a reference policy (or model). This quantity, RG^d, is constant with respect to the trainable parameters θ (because it depends on the reference model). However, it plays an important role in shaping how θ is learned for y_w^t and y_l^t.

Target Gap The difference $\beta\,\mathrm{RG}^t - \mathrm{RG}^d$ appears inside the loss function. Because RG^d is subtracted from $\beta\,\mathrm{RG}^t$, it effectively sets a target "goal" in log probabilities that the model π_θ should achieve between the winning (safe) candidate y_w^t and the losing (unsafe) candidate y_l^t.

Penalty for Reward Signal The term $\beta\,\mathrm{RG}^t - \mathrm{RG}^d$ penalizes deviations of $\beta\,\mathrm{RG}^t$ from RG^d. Intuitively, if $\beta\,\mathrm{RG}^t$ is not aligned with RG^d, the loss increases, thus signaling the training process that π_θ is not matching the reference gap.

Alignment with Reference Behavior Because RG^d comes from π_{ref}, one can interpret it as how strongly the reference policy prefers its "winning" candidate y_w^d over its "losing" candidate y_l^d. By forcing RG^t from π_θ to approximate RG^d, the training encourages π_θ to mimic or at least stay consistent with that preference structure, though for potentially different x^t and y_w^t, y_l^t.

Effect on y_w^t and y_l^t Learning For y_w^t: If the reference gap indicates a high preference for a corresponding "winning" candidate y_w^d, then during training, the model sees a stronger incentive to increase $\log \pi_\theta(y_w^t \mid x^t)$ (since that helps match the overall gap).

For y_l^t: The model similarly sees a signal to decrease $\log \pi_\theta(y_l^t \mid x^t)$ (or at least not let it grow too large), in order to keep the difference consistent with RG^d.

In essence, RG^d provides a reference or target difference in log probabilities that the model π_θ tries to match between y_w^t and y_l^t. Although it does not directly update θ (because it is constant with respect to θ), it influences the loss landscape and hence indirectly guides how $\log \pi_\theta(y_w^t \mid x^t)$ and $\log \pi_\theta(y_l^t \mid x^t)$ are learned.

C Gradient Analysis of MPO

C.1 Deriving the Gradient of MPO

Below is a step-by-step derivation of the gradient of the loss function.

$$\mathcal{L}(\pi_\theta) = \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\big\|\beta\,\mathrm{RG}^t - \mathrm{RG}^d\big\|^2\Big], \quad (12)$$

where

$$\mathrm{RG}^t = \frac{1}{|y_w^t|} \log \pi_\theta(y_w^t \mid x^t) - \frac{1}{|y_l^t|} \log \pi_\theta(y_l^t \mid x^t), \quad (13)$$

and

$$\mathrm{RG}^d = \frac{1}{|y_w^d|} \log \pi_{\text{ref}}(y_w^d \mid x^d) - \frac{1}{|y_l^d|} \log \pi_{\text{ref}}(y_l^d \mid x^d). \quad (14)$$

Note that RG^d does not depend on θ, whereas RG^t depends on θ through $\log \pi_\theta(\cdot)$.

Rewrite the Loss Function Define the per-sample loss (ignoring the expectation for a moment) as

$$\ell(\theta) \;=\; \left\|\beta\,\mathrm{RG}^t(\theta) \;-\; \mathrm{RG}^d\right\|^2. \tag{15}$$

For the gradient derivation, we focus on $\ell(\theta)$. The overall gradient will then be its expectation w.r.t. the data distribution $\mathcal{D}$.

Introduce an Intermediate Variable Let

$$z(\theta) \;=\; \beta\,\mathrm{RG}^t(\theta) \;-\; \mathrm{RG}^d. \tag{16}$$

Hence

$$\ell(\theta) \;=\; \left\|z(\theta)\right\|^2. \tag{17}$$

If RG^t is a scalar, $\|z(\theta)\|^2 = z(\theta)^2$. (For the vector case, one may treat each component in the same way.)

Apply the Chain Rule to $\ell(\theta)$ We have

$$\ell(\theta) \;=\; \left\|z(\theta)\right\|^2. \tag{18}$$

Taking the gradient w.r.t. θ,

$$\nabla_\theta\,\ell(\theta) \;=\; \nabla_\theta\left\|z(\theta)\right\|^2 \tag{19}$$

$$\;=\; 2\,z(\theta)\,\nabla_\theta\,z(\theta). \tag{20}$$

Recalling

$$z(\theta) \;=\; \beta\,\mathrm{RG}^t(\theta) \;-\; \mathrm{RG}^d, \tag{21}$$

and that RG^d is a constant w.r.t. θ, we get:

$$\nabla_\theta\,z(\theta) \;=\; \beta\,\nabla_\theta\,\mathrm{RG}^t(\theta). \tag{22}$$

Hence,

$$\nabla_\theta\,\ell(\theta) \;=\; 2\left[\beta\,\mathrm{RG}^t(\theta) - \mathrm{RG}^d\right]\beta\,\nabla_\theta\,\mathrm{RG}^t(\theta). \tag{23}$$

Compute $\nabla_\theta\,\mathrm{RG}^t(\theta)$ By definition,

$$\mathrm{RG}^t(\theta) \;=\; \frac{1}{|y_w^t|}\,\log \pi_\theta\big(y_w^t \mid x^t\big) \;-\; \frac{1}{|y_l^t|}\,\log \pi_\theta\big(y_l^t \mid x^t\big). \tag{24}$$

Hence,

$$\begin{aligned}\nabla_\theta\,\mathrm{RG}^t(\theta) = \frac{1}{|y_w^t|}\,\nabla_\theta\,\log \pi_\theta\big(y_w^t \mid x^t\big)\\ -\frac{1}{|y_l^t|}\,\nabla_\theta\,\log \pi_\theta\big(y_l^t \mid x^t\big).\end{aligned} \tag{25}$$

Combine the Results Putting it all together,

$$\nabla_\theta\,\ell(\theta) \;=\; 2\left[\beta\,\mathrm{RG}^t(\theta) - \mathrm{RG}^d\right]$$

$$\beta\left(\frac{1}{|y_w^t|}\,\nabla_\theta\,\log \pi_\theta(y_w^t \mid x^t) \;-\; \frac{1}{|y_l^t|}\,\nabla_\theta\,\log \pi_\theta(y_l^t \mid x^t)\right). \tag{26}$$

The Full Gradient of $\mathcal{L}(\theta)$ Recall the original loss is the *expectation* of $\ell(\theta)$ over samples $(x, y_w, y_l) \sim \mathcal{D}$. Therefore,

$$\nabla_\theta\,\mathcal{L}(\theta) \;=\; 2\beta\,\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Bigg[\left(\beta\,\mathrm{RG}^t(\theta) - \mathrm{RG}^d\right)$$

$$\left(\frac{1}{|y_w^t|}\,\nabla_\theta\,\log \pi_\theta(y_w^t \mid x^t) \;-\; \frac{1}{|y_l^t|}\,\nabla_\theta\,\log \pi_\theta(y_l^t \mid x^t)\right)\Bigg]. \tag{27}$$

This completes the derivation of the gradient w.r.t. the parameters θ.

C.2 Analysis

Here we explain how RG^d influences the model's updates for y_w^t and y_l^t from the gradient view:

Shifts the Gradient Magnitude and Direction The difference $(\beta\,\mathrm{RG}^t(\theta) - \mathrm{RG}^d)$ multiplies the gradient terms that involve $\log \pi_\theta(y_w^t \mid x^t)$ and $\log \pi_\theta(y_l^t \mid x^t)$. If $\beta\,\mathrm{RG}^t$ is smaller than RG^d, then the difference is negative, which encourages the model to increase RG^t (e.g., by increasing the probability of y_w^t or decreasing the probability of y_l^t) so that it moves toward or surpasses RG^d. If $\beta\,\mathrm{RG}^t$ is larger than RG^d, the difference is positive, so the model is nudged to preserve or even enlarge its current gap, reinforcing the discrimination it has already learned between y_w^t and y_l^t.

Controls the Drive to Differentiate y_w^t and y_l^t Because RG^t involves $\log \pi_\theta(y_w^t)$ and $\log \pi_\theta(y_l^t)$, the difference term $(\beta\,\mathrm{RG}^t(\theta) - \mathrm{RG}^d)$ directly scales how strongly the model updates its parameters to favor y_w^t over y_l^t. A larger RG^d essentially raises the "bar" the model is trying to clear; a smaller RG^d lowers it.

D Training Data

To ensure that our multilingual preference data generation process remains on-policy, we adopt a structured approach based on well-established principles in LLM post-training (Dubey et al., 2024). Specifically, for each English harmful prompt in the PKU-SafeRLHF dataset (Ji et al., 2024), we first feed it to the model to generate a refusal response, which serves as the preferred response. We then pair this generated refusal with the original dispreferred response from the dataset, forming a preference pair. This ensures that the optimization process remains aligned with the model's actual behavior, avoiding potential inconsistencies that arise from using static preference data (Yuan et al.,

2024; Chen et al., 2024; Rosset et al., 2024; Wu et al., 2024b; Guo et al., 2024; Zhou et al., 2025).

To extend this preference data to multiple languages, we then translate both the harmful prompts and their paired responses using the Google Translate API. This approach allows us to create multilingual preference data while preserving the preference structure of the original dataset.

E Benchmark

We comprehensively measure the efficacy of our MPO on various multilingual safety benchmarks.

- **MultiJail** (Deng et al., 2024): It carefully gather 315 English harmful queries and manually translate them by native speakers into 9 non-English languages, ranging from high-resource to low-resource.

- **AdvBench-X** (Yong et al., 2023): AdvBench is a set of 500 harmful behaviors formulated as instructions. These behaviors range over the same themes as the harmful strings setting, but the adversary's goal is to find a single attack string that will cause the model to generate any response that attempts to comply with the instruction, and to do so over as many harmful behaviors as possible. The original English version is also translated manually into target languages of different resource levels.

- **CSRT** (Yoo et al., 2024): It synthesizes code-switching red-teaming queries, combining up to 10 languages, and investigate the safety and multilingual understanding of LLMs.

We evaluate multilingual safety alignment using the Attack Success Rate (ASR), following the evaluation pipeline proposed by Deng et al. (2024), with GPT-4o as the judgment model. The evaluation process consists of the following steps: (1) Translation to English: Since safety alignment performance needs to be assessed across multiple languages, we first translate the model-generated responses from the target language into English using GPT-4o to ensure consistent evaluation. (2) Three-Class Classification: GPT-4o then classifies each response into one of the following categories: Safe (meaningful refusal), Unsafe or Irrelevant. (3) Attack Success Calculation: Responses classified as unsafe or irrelevant are both considered unsuccessful refusals and thus counted as successful attacks when calculating ASR. Only safe refusals are considered failed attacks, contributing to a lower ASR (better safety performance). Importantly, as shown in Figure 2 of Deng et al. (2024), the evaluation pipeline combining translation and GPT-4 classification achieves strong agreement with human annotations, with a Cohen's kappa score of 0.86, indicating a high level of alignment. This validates the reliability and soundness of the evaluation protocol.

It is essential to highlight that the languages targeted for enhancement, as mentioned above, are all within the capability range of GPT-4o, especially given that its official model card (OpenAI, 2024a) emphasizes support for low-resource languages such as Swahili (Sw) and Bengali (Bn). This underscores the validity and reliability of the evaluation approach.

F Baseline Methods

We compare MPO with other preference optimization methods listed in Table 7. IPO (Azar et al., 2024) is a theoretically grounded approach that avoids DPO's assumption that pairwise preferences can be replaced with pointwise rewards. rDPO (Chowdhury et al., 2024) mitigates the impact of noise on average, making policies trained with this method more robust. CPO (Xu et al., 2024a) leverages sequence likelihood as a reward and trains jointly with an SFT objective. KTO (Ethayarajh et al., 2024) learns from non-paired preference data, while ORPO (Hong et al., 2024b) introduces a reference-model-free odds ratio term to directly contrast winning and losing responses with the policy model, training it alongside the SFT objective. R-DPO (Park et al., 2024) modifies DPO by incorporating an additional regularization term to prevent length exploitation. Finally, SimPO (Meng et al., 2024) normalizes rewards based on response length and enforces a target reward margin, ensuring that the reward difference between winning and losing responses meets a predefined threshold.

G Implementation Details

All training experiments are conducted on eight A100 GPUs using the LLaMA-Factory repository (Zheng et al., 2024b). And our MPO is also implement based on this repo. For distributed training, we leverage the DeepSpeed (Rasley et al., 2020) framework with ZeRo-2 optimization. Initially, we perform preliminary experiments to determine optimal batch sizes from [8, 16, 32] and training epochs from [1, 2, 3]. We observe that a batch size

Method	Objective	Hyperparameter
DPO (Rafailov et al., 2023)	$-\log\sigma\left(\beta\log\frac{\pi_\theta(y_w\|x)}{\pi_{\text{ref}}(y_w\|x)}-\beta\log\frac{\pi_\theta(y_l\|x)}{\pi_{\text{ref}}(y_l\|x)}\right)$	$\beta\in[0.01,0.05,0.1]$
IPO (Azar et al., 2024)	$\left(\log\frac{\pi_\theta(y_w\|x)}{\pi_{\text{ref}}(y_w\|x)}-\log\frac{\pi_\theta(y_l\|x)}{\pi_{\text{ref}}(y_l\|x)}-\frac{1}{2\tau}\right)^2$	$\tau\in[0.01,0.1,0.5,1.0]$
rDPO (Chowdhury et al., 2024)	$\frac{(1-\epsilon)\mathcal{L}(\theta,x,y_w,y_l)-\epsilon\mathcal{L}(\theta,x,y_l,y_w)}{1-2\epsilon}$ $\mathcal{L}(\theta,x,y_l,y_w)=-\log\sigma\left(\beta\log\frac{\pi_\theta(y_w\|x)}{\pi_{\text{ref}}(y_w\|x)}-\beta\log\frac{\pi_\theta(y_l\|x)}{\pi_{\text{ref}}(y_l\|x)}\right)$	$\epsilon\in[01,0.5]$ $\beta\in[0.01,0.05,0.1]$
CPO (Xu et al., 2024a)	$-\log\sigma\left(\beta\log\pi_\theta(y_w\|x)-\beta\log\pi_\theta(y_l\|x)\right)-\lambda\log\pi_\theta(y_w\|x)$	$\lambda=1.0,\ \beta\in[0.01,0.05,0.1]$
KTO (Ethayarajh et al., 2024)	$-\lambda_w\sigma\left(\beta\log\frac{\pi_\theta(y_w\|x)}{\pi_{\text{ref}}(y_w\|x)}-z_{\text{ref}}\right)+\lambda_l\sigma\left(z_{\text{ref}}-\beta\log\frac{\pi_\theta(y_l\|x)}{\pi_{\text{ref}}(y_l\|x)}\right),$ where $z_{\text{ref}}=\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\beta\text{KL}\left(\pi_\theta(y\|x)\|\pi_{\text{ref}}(y\|x)\right)\right]$	$\lambda_l=\lambda_w=1.0$ $\beta\in[0.01,0.1,1.0]$
ORPO (Hong et al., 2024b)	$-\log p_\theta(y_w\|x)-\lambda\log\sigma\left(\log\frac{p_\theta(y_w\|x)}{1-p_\theta(y_w\|x)}-\log\frac{p_\theta(y_l\|x)}{1-p_\theta(y_l\|x)}\right),$ where $p_\theta(y\|x)=\exp\left(\frac{1}{\|y\|}\log\pi_\theta(y\|x)\right)$	$\lambda\in[0.01,0.1,1.0]$
R-DPO (Park et al., 2024)	$-\log\sigma\left(\beta\log\frac{\pi_\theta(y_w\|x)}{\pi_{\text{ref}}(y_w\|x)}-\beta\log\frac{\pi_\theta(y_l\|x)}{\pi_{\text{ref}}(y_l\|x)}+(\alpha\|y_w\|-\alpha\|y_l\|)\right)$	$\alpha\in[0.05,0.1,0.5,1.0]$ $\beta\in[0.01,0.05,0.1]$
SimPO (Meng et al., 2024)	$-\log\sigma\left(\frac{\beta}{\|y_w\|}\log\pi_\theta(y_w\|x)-\frac{\beta}{\|y_l\|}\log\pi_\theta(y_l\|x)-\gamma\right)$	$\beta\in[2.0,2.5]$ $\gamma\in[1.0,1.2,1.4,1.6]$

Table 7: Detailed optimization objectives of current preference learning methods. We carefully tune their specific hyperparameters and list the search space in the right column.

	Learning Rate	Epoch	β
LLaMA-3.1-8B-Instruct	6e-7	2	1.0
Gemma-2-9b-it	4e-7	2	1.5
Qwen2.5-7B-Instruct	6e-7	2	1.5

Table 8: The hyperparameters in our proposed MPO used for all three backbones.

of 8 consistently yields the best performance across all methods, while the optimal number of training epochs varies by method. All models on all three backbones are trained with a maximum sequence length of 2048, and we employ a cosine learning rate schedule with a 10% warmup phase.

To further refine performance, we extensively tune key hyperparameters for all baselines, including the learning rate, training epochs, and method-specific parameters. The learning rate is searched within [3e-7, 4e-7, 5e-7, 6e-7, 1e-6], while training epochs are explored in [1, 2, 3]. Method-specific hyperparameter search spaces are detailed in Table 7. For MPO, β is searched in [1.0, 1.5, 2.0] and we find that 1.0 or 1.5 always exhibit the best results across all three backbones. Table 8 shows MPO's hyperparameters used under each backbone.

H Additional Experimental Results

H.1 Results on Qwen2.5

Table 9 demonstrates the performance comparison of MPO and baselines based on Qwen2.5-7B-Instruct. we have drawn the following key insights:

MPO still exhibits robust and consistent performance across various benchmarks and maintain multilingual utility. It consistently surpasses all preference learning methods, highlighting its outstanding safety alignment capabilities and scalability. Tables 10, 11 and 12 presents MPO maintains multilingual utility on MT-Bench, M-MMLU and MGSM, respectively.

Multilingual safety alignment depends on foundational abilities The improvement of multilingual safety performance relies on the foundational multilingual capabilities of the backbone model. Results on Qwen2.5 show that while MPO still achieves significant gains compared to the original model and baselines, its absolute performance lags behind other two backbones, especially for low-resource languages. This disparity arises from Qwen2.5's weaker foundational abilities in these languages. More Specifically, as shown in Table 10, Qwen2.5 exhibits weak instruction-following ability in Bn and Sw, frequently generating outputs unrelated to the input. In our evaluation, such outputs are classified as unsafe.

H.2 Evaluation on Multilingual Utility

Evaluation Settings We conduct a comprehensive evaluation of MPO's impact on multilingual utility across the following benchmarks.

- **MT-Bench** (Zheng et al., 2023): The dataset is designed for open-ended generation to

	MultiJail							AdvBench-X								CSRT
	En	**Zh**	**Ko**	**Ar**	**Bn**	**Sw**	**AVG.**	**En**	**Zh**	**Jp**	**Ko**	**Ar**	**Bn**	**Sw**	**AVG.**	**-**
Qwen2.5	12.70	10.16	15.87	15.87	73.02	98.10	37.62	1.15	1.35	5.96	5.38	6.35	57.58	99.04	25.26	34.60
SFT	10.79	10.79	13.02	13.02	64.76	99.05	35.24	1.35	2.12	5.38	3.46	5.38	48.18	98.46	23.48	34.92
DPO	11.43	10.48	12.38	13.33	69.84	98.73	36.03	1.54	1.35	5.77	3.85	4.81	51.82	98.08	23.89	37.46
IPO	11.43	8.89	13.65	13.02	68.25	99.05	35.72	**0.96**	1.92	5.96	5.19	5.77	53.93	97.70	24.49	38.10
rDPO	9.84	8.25	15.24	14.92	70.16	98.73	36.19	1.73	1.92	4.23	4.04	6.92	52.78	96.93	24.08	35.56
CPO	13.33	7.94	12.38	13.33	58.92	98.73	34.11	1.15	2.31	5.96	4.23	6.35	50.48	99.04	24.22	34.92
KTO	10.16	9.52	13.65	13.97	66.67	99.05	35.50	2.50	1.54	5.38	4.04	6.73	53.93	98.08	24.60	39.81
ORPO	10.16	10.16	16.19	13.97	67.62	99.37	36.25	2.31	2.50	5.77	3.27	4.62	48.56	98.46	23.64	30.16
R-DPO	10.79	7.62	9.84	14.29	58.41	98.10	33.18	1.73	2.12	5.58	4.62	5.58	56.24	98.85	24.96	38.41
SimPO	11.75	9.52	14.60	13.97	70.79	98.41	36.51	1.35	1.92	5.38	4.23	5.96	49.33	98.85	23.86	31.43
MPO (Ours)	**7.30**	**6.67**	**8.89**	**13.02**	**53.65**	**92.38**	**30.32**	1.92	**0.96**	**3.27**	**2.50**	**3.65**	**30.33**	**85.03**	**18.24**	**26.35**
MPO - En Align	9.52	13.65	13.33	13.97	64.44	98.73	35.61	2.12	3.46	3.27	2.88	5.00	46.64	97.50	22.98	27.94

Table 9: Detailed results of Qwen2.5-7B-Instruct on three multilingual safety benchmarks are presented. The evaluation metric used is the Attack Success Rate (ASR), where lower values indicate better performance. The best results achieved by our method and baselines are highlighted in bold, while the second-best results are underlined.

	MT-Bench							
	En	**Zh**	**Jp**	**Ko**	**Ar**	**Bn**	**Sw**	**AVG.**
LLaMA-3.1	7.31	5.38	4.88	5.22	5.43	3.98	3.98	5.17
SFT	7.31	5.56	4.84	4.94	5.09	4.25	3.72	5.10
DPO	7.44	5.66	5.03	5.49	4.89	4.76	4.16	5.35
IPO	7.31	5.42	4.89	5.12	5.23	4.39	4.08	5.26
rDPO	7.31	5.81	5.31	5.16	5.43	4.44	4.21	5.38
CPO	7.45	5.59	4.98	4.93	5.04	4.16	3.86	5.14
KTO	7.33	5.55	5.02	5.11	5.05	4.39	4.01	5.24
ORPO	7.39	5.41	4.73	5.01	5.36	4.24	3.72	5.12
R-DPO	7.30	5.63	5.21	5.45	5.48	4.80	4.11	5.43
SimPO	7.48	5.48	5.54	5.21	5.59	4.01	4.11	5.35
MPO	7.25	5.32	5.26	5.44	5.38	4.11	4.01	5.25
Gemma-2	7.71	7.07	6.84	6.81	7.06	5.66	6.15	6.76
SFT	7.72	6.86	6.31	6.38	7.08	5.16	5.84	6.48
DPO	7.79	6.88	7.06	6.86	6.98	6.26	6.20	6.86
IPO	7.61	6.86	6.95	6.86	7.07	5.88	6.17	6.77
rDPO	7.57	6.82	6.93	6.57	6.98	6.03	6.23	6.73
CPO	7.73	6.64	6.62	6.56	7.01	5.33	5.98	6.55
KTO	7.63	6.87	7.00	6.69	6.88	6.06	6.04	6.74
ORPO	7.71	6.86	6.78	6.49	7.09	5.22	6.07	6.60
R-DPO	7.77	7.11	7.09	7.33	6.87	5.33	6.11	6.80
SimPO	7.47	6.89	7.00	6.78	6.95	5.84	6.16	6.73
MPO	7.83	6.81	7.07	6.88	7.05	5.78	6.16	6.80
Qwen2.5	7.77	7.36	6.43	6.21	6.60	4.49	2.37	5.89
SFT	7.49	7.14	6.73	6.50	6.46	4.66	2.08	5.87
DPO	7.68	6.99	6.79	6.61	6.50	4.71	2.24	5.93
IPO	7.66	7.24	6.86	6.29	6.61	4.68	1.98	5.90
rDPO	7.78	7.01	6.58	6.38	6.59	4.48	2.10	5.85
CPO	7.61	7.16	6.82	6.33	6.57	4.54	2.12	5.88
KTO	7.77	7.03	6.82	6.60	6.64	4.71	2.18	5.96
ORPO	7.52	7.18	6.58	6.43	6.78	4.53	2.09	5.87
R-DPO	7.61	7.23	6.90	6.33	6.56	4.87	2.20	5.96
SimPO	7.57	7.07	6.73	6.29	6.68	4.77	2.08	5.88
MPO	7.77	7.39	6.71	6.19	6.56	4.43	2.28	5.90

Table 10: Results on MT-Bench across three backbones.

	M-MMLU							
	En	**Zh**	**Jp**	**Ko**	**Ar**	**Bn**	**Sw**	**AVG.**
LLaMA-3.1	67.70	51.30	47.90	43.30	47.60	41.40	40.60	48.54
SFT	67.30	50.90	47.70	47.00	42.60	40.60	39.20	47.90
DPO	67.10	51.50	48.40	47.30	41.80	39.80	38.20	47.73
IPO	67.30	51.60	48.30	47.90	43.00	41.30	39.60	48.43
rDPO	67.20	50.80	47.90	47.30	42.80	40.00	40.50	48.07
CPO	67.40	51.60	47.60	47.50	42.50	40.00	39.80	48.06
KTO	67.10	51.40	48.50	47.60	41.90	40.40	40.40	48.19
ORPO	67.30	51.00	48.10	47.30	42.10	41.30	38.90	48.00
R-DPO	66.90	51.50	48.10	47.70	43.30	40.70	40.20	48.34
SimPO	66.70	51.40	47.70	47.80	43.00	40.90	40.40	48.27
MPO	67.10	50.70	48.40	42.40	47.70	40.10	38.70	47.87
Gemma-2	73.40	61.20	59.40	53.80	59.10	49.90	52.40	58.45
SFT	61.10	50.70	73.30	59.40	59.40	55.40	52.50	58.83
DPO	61.20	49.80	73.40	58.70	59.20	53.90	52.20	58.34
IPO	61.30	49.70	73.30	59.30	59.50	53.60	52.40	58.44
rDPO	61.00	50.20	73.30	59.10	59.40	54.20	52.50	58.53
CPO	61.40	50.80	73.30	59.30	59.60	55.10	52.60	58.87
KTO	61.40	49.90	73.40	59.20	59.50	53.90	52.40	58.53
ORPO	61.70	50.40	73.20	59.30	59.70	55.20	52.90	58.91
R-DPO	60.80	49.00	73.30	59.10	59.20	54.00	51.40	58.11
SimPO	61.30	49.90	73.30	59.20	59.40	53.90	52.30	58.47
MPO	73.40	61.20	58.90	54.00	59.50	49.80	52.10	58.41
Qwen2.5	72.50	63.90	57.70	49.70	56.60	43.20	31.50	53.59
SFT	64.10	43.10	72.60	56.70	57.80	49.30	31.30	53.56
DPO	64.00	43.10	72.50	56.70	57.70	49.50	31.40	53.56
IPO	64.20	43.10	72.20	57.00	57.30	49.70	31.40	53.56
rDPO	64.00	43.40	72.40	56.60	57.70	49.70	31.40	53.60
CPO	64.40	42.60	72.70	57.00	57.90	49.10	31.60	53.61
KTO	64.00	43.20	72.50	56.70	57.70	49.80	31.40	53.61
ORPO	64.50	43.00	72.70	57.10	57.80	50.40	31.50	53.86
R-DPO	64.20	43.30	72.50	56.50	57.60	49.60	31.40	53.59
SimPO	64.10	43.10	72.20	56.80	57.70	50.00	31.40	53.61
MPO	72.20	64.50	57.30	50.50	57.10	43.20	31.60	53.77

Table 11: Results on M-MMLU across three backbones.

	MGSM				
	En	**Zh**	**Bn**	**Sw**	**AVG.**
LLaMA-3.1	88.00	67.20	12.40	40.80	52.10
SFT	86.80	69.60	13.60	54.00	56.00
DPO	85.60	68.80	15.60	45.20	53.80
IPO	85.60	68.80	16.00	46.80	54.30
rDPO	86.80	71.20	13.60	39.60	52.80
CPO	88.80	73.60	11.20	43.20	54.20
KTO	84.80	67.60	16.80	48.40	54.40
ORPO	86.80	69.60	14.80	55.20	56.60
R-DPO	86.80	71.20	13.60	39.60	52.80
SimPO	86.00	72.40	11.60	46.40	54.10
MPO	88.00	68.40	12.00	53.60	55.50
Gemma-2	90.00	77.60	66.00	75.20	77.20
SFT	89.60	79.20	43.60	63.60	69.00
DPO	89.20	78.80	67.60	73.60	77.30
IPO	90.00	78.40	67.20	75.20	77.70
rDPO	90.40	78.00	65.60	76.80	77.70
CPO	90.40	79.20	46.00	68.00	70.90
KTO	90.00	77.60	67.60	75.20	77.60
ORPO	90.40	80.00	49.60	65.20	71.30
R-DPO	90.00	77.60	75.20	75.60	79.60
SimPO	90.00	76.80	67.20	75.20	77.30
MPO	90.80	80.40	70.00	74.00	78.80
Qwen2.5	87.20	82.00	35.20	6.40	52.70
SFT	87.60	82.00	35.60	8.40	53.40
DPO	87.60	82.00	38.00	7.20	53.70
IPO	88.00	84.40	36.40	7.60	54.10
rDPO	88.00	82.40	36.00	8.00	53.60
CPO	88.80	82.40	34.00	9.20	53.60
KTO	87.60	82.40	36.40	7.60	53.50
ORPO	88.40	82.40	24.00	6.80	50.40
R-DPO	87.20	82.40	36.40	7.20	53.30
SimPO	88.40	84.00	38.00	9.20	54.90
MPO	88.40	82.80	34.40	6.80	53.10

Table 12: Results on MGSM across three backbones.

evaluate a model's ability to follow multi-turn instructions. In our experimental setup, this benchmark covers English (En), Chinese (Zh), Arabic (Ar), Japanese (Jp), Korean (Ko), Swahili (Sw) and Bengali (Bn). We collect data in English[1], Japanese[2], Korean[3], and Arabic[4] from huggingface, and Chinese[5] from github. In addition, we use GPT-4o to translate the English data into Swahili and Bengali, and performed manual proofreading to ensure correctness. The evaluation follows the **LLM-as-a-judge** approach, where GPT-4o is prompted to assign a score directly to a single response on a scale of 1 to 10. It is essential to highlight that the languages targeted for enhancement, as mentioned above, are all within the capability range of GPT-4o, especially given that its official model card (OpenAI, 2024a) emphasizes support for low-resource languages such as Swahili (Sw) and Bengali (Bn). This underscores the validity and reliability of the evaluation approach.

- **M-MMLU** (Hendrycks et al., 2021):[6] The MMLU is a widely recognized benchmark of general knowledge attained by AI models. It covers a broad range of topics from 57 different categories, covering elementary-level knowledge up to advanced professional subjects like law, physics, history, and computer science. OpenAI translated the MMLU's test set into 14 languages using professional human translators. Relying on human translators for this evaluation increases confidence in the accuracy of the translations, especially for low-resource languages. In our experimental setup, we adopt the 5-shot evaluation and this benchmark covers English (En), Chinese (Zh), Japanese (Jp), Arabic (Ar), Korean (Ko), Swahili (Sw) and Bengali (Bn).

- **MGSM** (Shi et al., 2023):[7] Multilingual Grade School Math Benchmark (MGSM) is

[1] https://huggingface.co/datasets/HuggingFaceH4/mt_bench_prompts
[2] https://huggingface.co/datasets/shi3z/MTbenchJapanese
[3] https://huggingface.co/datasets/StudentLLM/Korean_MT-Bench_questions
[4] https://huggingface.co/spaces/QCRI/mt-bench-ar/tree/main/data/mt_bench_ar
[5] https://github.com/HIT-SCIR/huozi
[6] https://huggingface.co/datasets/openai/MMMLU
[7] https://huggingface.co/datasets/juletxara/mgsm

a benchmark of grade-school math problems. The same 250 problems from GSM8K (Cobbe et al., 2021) are each translated via human annotators in 10 languages. The dataset was created to support the task of question answering on basic mathematical problems that require multi-step reasoning. In our experimental setup, we performance evaluation via 0-shot CoT (Wei et al., 2022) and this benchmark covers English (En), Chinese (Zh), Swahili (Sw) and Bengali (Bn).

Detailed Results Here, we demonstrate the detailed results and the comparison with baseline methods. For MT-Bench, results on LLaMA-3.1-8B-Instruct, Gemma-2-9B-it and Qwen2.5-7B-Instruct are shown in Table 10. For M-MMLU, results are shown in Table 11. For MGSM, results are shown in Table 12. The results across three backbones show that MPO consistently maintains the general utility of both the dominant and target languages, demonstrating its effectiveness in achieving multilingual safety alignment without compromising the model's multilingual utility.

Current preference learning methods typically incorporate a KL constraint during training to prevent the model from deviating too far from its original state, ensuring that multilingual general utility is well preserved. As a result, these methods maintain multilingual capabilities comparable to the original model, even after alignment.

Under the same achievement of preserving multilingual general utility, MPO achieves significantly superior multilingual safety performance compared to these methods. By leveraging reward gap minimization with the dominant language as a high-quality supervision signal, MPO effectively transfers safety alignment across languages without degrading the model's overall linguistic competence. This highlights its advantage in balancing multilingual safety and utility, making it a more effective approach for multilingual safety alignment in LLMs. And the comparison of KL constraint and the representation constraint used in MPO is further discussed in Appendix H.3.

H.3 Ablation Study

Fixed Constants as the Supervision Signal Table 13 presents the detailed results of multilingual safety performance and general utility of the model when replacing the dominant language reward gap with a fixed value ranging from 0.1 to 20. Notably, 1.58 corresponds to the average reward gap of dominant language samples in the training set. As the constant increases, multilingual safety performance steadily improves, even exceeding the performance of models aligned using the actual dominant language reward gap. However, this improvement comes at a significant cost to multilingual general utility, as excessive alignment strength induces substantial parameter shifts, leading to model collapse despite the application of a retention constraint. Additionally, setting the constant to 1.58 yields only limited improvements, suggesting that fine-grained supervision at the sample level is superior to coarse-grained dataset-level alignment.

Reward Gap of Other Languages as the Supervision Signal Table 5 further demonstrates that using the reward gap of a target language as the alignment objective fails to yield meaningful safety improvements. When selecting the second-best safety-performing language (Arabic) or even low-resource languages (Swahili, Bengali), no effective multilingual safety enhancement is observed. This reinforces that the dominant language's reward gap provides a more reliable and high-quality alignment supervision signal.

Comparison with Cross-Lingual Transfer Methods Cross-lingual transfer methods posit that skills acquired in one source language can be effectively transferred to other languages (Huang et al., 2023; Ranaldi et al., 2023; Qin et al., 2023; Etxaniz et al., 2024). This has been achieved through two main approaches: aligning multilingual representations with the activation space of LLMs: CLA (Li et al., 2024a) and LENS (Zhao et al., 2024a), or distilling knowledge from the dominant language: SDRRL (Zhang et al., 2024b). The details of these recent advancements are as follows:

- **CLA**: It aligns internal sentence representations across languages through multilingual contrastive learning and ensures output alignment by adhering to cross-lingual instructions in the target language.

- **LENS**: It enhances multilingual capabilities by leveraging LLMs' internal language representation spaces. LENS operates on two subspaces: the language-agnostic subspace, where it aligns target languages with the central language to inherit strong semantic representations, and the language-specific sub-

	MultiJail							MT-Bench							
	En	**Zh**	**Ko**	**Ar**	**Bn**	**Sw**	**AVG.**	**En**	**Zh**	**Jp**	**Ko**	**Ar**	**Bn**	**Sw**	**AVG.**
LLaMA-3.1	14.60	20.32	52.38	16.83	49.52	37.78	31.91	7.31	5.38	4.88	5.22	5.43	3.98	3.98	5.17
Constant 0.1	25.08	46.67	61.90	63.38	59.37	85.08	56.91	7.13	5.18	4.72	4.47	4.67	4.16	3.39	4.82
Constant 0.5	10.16	8.89	26.98	14.29	21.59	45.08	21.17	**7.34**	**5.49**	4.77	5.23	5.04	4.08	3.68	5.09
Constant 1.0	2.86	1.59	11.11	1.90	7.30	18.41	7.20	7.12	5.26	4.71	5.09	5.08	3.97	3.71	4.99
Constant 1.58	2.86	1.27	6.03	0.32	6.35	13.97	5.13	7.26	5.16	5.16	5.18	5.02	**4.27**	**4.14**	5.17
Constant 2.0	**0.32**	**0.00**	**0.63**	**0.00**	**5.71**	6.03	**2.12**	7.04	4.43	4.01	5.26	3.43	3.61	3.54	4.47
Constant 5.0	**0.32**	**0.00**	**0.63**	1.90	7.30	9.21	3.23	**7.34**	4.81	4.53	5.22	3.88	4.00	3.02	4.69
Constant 10.0	**0.32**	0.32	0.95	0.63	6.35	6.03	2.43	6.88	4.16	3.64	4.20	3.06	3.63	3.58	4.16
Constant 20.0	**0.32**	0.32	**0.63**	0.63	14.60	**3.81**	3.39	7.21	4.94	4.31	5.16	4.03	3.26	3.61	4.65
MPO (Ours)	2.22	0.95	4.76	1.90	12.38	10.79	5.98	7.25	5.32	**5.26**	**5.44**	**5.38**	4.11	4.01	**5.25**

Table 13: Results of the multilingual safety performance and general utility of the model when replacing the dominant language reward gap with a fixed value ranging from 0.1 to 20. The best results are highlighted in bold.

	MultiJail						
	En	**Zh**	**Ko**	**Ar**	**Bn**	**Sw**	**AVG.**
LLaMA-3.1	14.60	20.32	52.38	16.83	49.52	37.78	31.91
SDRRL	4.76	4.13	18.41	3.81	21.90	21.59	12.43
CLA	11.75	15.24	43.49	12.87	42.54	55.56	30.24
LENS	16.09	60.95	55.24	25.08	60.95	80.32	49.77
MPO	**2.22**	**0.95**	**4.76**	**1.90**	**12.38**	**10.79**	**5.98**

Table 14: Comparison with cross-lingual transfer method on MultiJail. The evaluation metric used is the Attack Success Rate (ASR), where lower values indicate better performance. The best results are highlighted in bold.

space, where it separates target and central languages to preserve linguistic specificity.

- **SDRRL**: It leverages self-distillation from resource-rich languages to effectively enhance multilingual performance through the use of self-distilled data.

Table 14 demonstrates that MPO consistently outperforms these methods, maintaining strong multilingual safety alignment. This further highlights the advantage of leveraging the dominant language's reward gap as a fine-grained supervision signal. Unlike MPO, which explicitly minimizes the reward gap difference between the dominant language and target languages, existing cross-lingual transfer approaches struggle with noisy preference signals and suboptimal knowledge transfer. Additionally, they often exhibit performance degradation in low-resource languages, where data scarcity amplifies alignment instability.

Effect of Constant Reward Gap from the Domiant Language We consider an alternative design of MPO—namely, computing the dominant language reward gap using the policy model instead of the reference model (MPO-Policy).

We implement MPO - Policy on all three backbones, LLaMA-3.1, Gemma-2 and Qwen2.5, where the dominant reward gap is computed using the policy model during training. We performed an equivalent hyperparameter search over learning rates [3e-7, 4e-7, 5e-7, 6e-7], training epochs [1, 2, 3], and values [1.0, 1.5, 2.0], and report the best-performing configuration. The results, shown in the Table 15, indicate that MPO - Policy consistently underperforms compared to our proposed method across all three backbones, both in terms of safety alignment (as measured by MultiJail, the lower score is better) and general capabilities (as measured by MT-Bench, the higher score is better). This empirically supports our choice to use the reference-based reward gap.

We further provide theoretical justification. When computing RG^d using the current model π_θ instead of the reference model π_{ref}, The gradient of this loss function becomes:

$$\nabla_\theta \mathcal{L}1 = 2\beta\, \mathbb{E}\mathcal{D}\Big[(\beta \cdot \mathrm{RG}^t - \mathrm{RG}^d)\cdot \Big(\Big(\frac{1}{|y_w^t|}\nabla_\theta \log \pi_\theta(y_w^t|x^t) - \frac{1}{|y_l^t|}\nabla_\theta \log \pi_\theta(y_l^t|x^t)\Big) - \Big(\frac{1}{|y_w^d|}\nabla_\theta \log \pi_\theta(y_w^d|x^d) - \frac{1}{|y_l^d|}\nabla_\theta \log \pi_\theta(y_l^d|x^d)\Big)\Big)\Big] \tag{28}$$

By comparing the gradients of the two formulations, Equation 27 and Equation 28, obtaining dominant reward gap from the policy model has the following drawbacks:

- **Optimization instability**: the direction of the gradient can fluctuate significantly as both sides of the loss are functions of θ.
- **Lack of anchoring**: without a stable reference, the loss can converge to trivial solutions

	MultiJail							MT-Bench							
	En	**Zh**	**Ko**	**Ar**	**Bn**	**Sw**	**AVG.**	**En**	**Zh**	**Jp**	**Ko**	**Ar**	**Bn**	**Sw**	**AVG.**
LLaMA-3.1	14.60	20.32	52.38	16.83	49.52	37.78	31.91	7.31	5.38	4.88	5.22	5.43	3.98	3.98	5.17
MPO	**2.22**	**0.95**	**4.76**	**1.90**	**12.38**	**10.79**	**5.98**	**7.25**	**5.32**	**5.26**	**5.44**	**5.38**	4.11	**4.01**	**5.25**
MPO - Policy	56.51	48.89	70.79	46.35	81.27	84.13	64.66	7.04	5.23	5.03	5.34	5.36	4.11	3.97	5.15
Gemma-2	2.54	9.52	14.61	4.13	20.32	14.60	10.95	7.71	7.07	6.84	6.81	7.06	5.66	6.15	6.76
MPO	**0.63**	**4.76**	**6.98**	**3.81**	**16.51**	**7.94**	**6.77**	**7.83**	**6.81**	**7.07**	**6.88**	**7.05**	**5.78**	**6.16**	**6.80**
MPO - Policy	1.59	5.08	7.30	5.40	17.78	10.83	7.99	7.81	6.77	6.73	6.59	6.84	5.71	6.01	6.64
Qwen-2.5	12.70	10.16	15.87	15.87	73.02	98.10	37.62	7.77	7.36	6.43	6.21	6.60	4.49	2.37	5.89
MPO	**7.30**	**6.67**	**8.89**	**13.02**	**53.65**	**92.38**	**30.32**	**7.77**	**7.39**	**6.71**	**6.19**	**6.56**	**4.43**	**2.28**	**5.90**
MPO - Policy	14.60	10.48	12.70	13.33	60.63	98.73	35.08	7.69	6.83	6.57	6.16	6.48	4.13	2.17	5.72

Table 15: Results of the multilingual safety performance and general utility of the model when replacing the computation of the dominant language reward gap with the policy model itself.

	MultiJail						
	En	**Zh**	**Ko**	**Ar**	**Bn**	**Sw**	**AVG.**
MAPO - DPO	5.40	3.49	15.87	3.17	27.71	42.86	16.42
MAPO - MPO	**2.46**	**1.59**	**3.17**	**3.17**	**9.52**	**8.25**	**4.69**
LIDR - DPO + NLL	**1.90**	6.03	18.10	19.37	80.00	81.27	34.45
LIDR - MPO	2.22	**2.86**	**1.90**	**6.67**	**8.89**	**11.12**	**5.61**

Table 16: Results of the multilingual safety performance on MultiJail. The evaluation metric used is the Attack Success Rate (ASR), where lower values indicate better performance. The best results are highlighted in bold.

where both RG^t and RG^d collapse toward zero, rather than aligning their structure.

In summary, our current design using reward gap from reference model not only improves training stability but also provides a clearer learning signal, enabling more reliable cross-lingual safety alignment. This design choice is well-justified, as the dominant language in the original model typically exhibits the strongest safety alignment

H.4 Impact of Data Source

Recent studies explore the use of LLMs themselves to generate multilingual preference data, rather than relying on external translation tools. Two notable approaches in this direction are MAPO (She et al., 2024) and LIDR (Yang et al., 2024b).

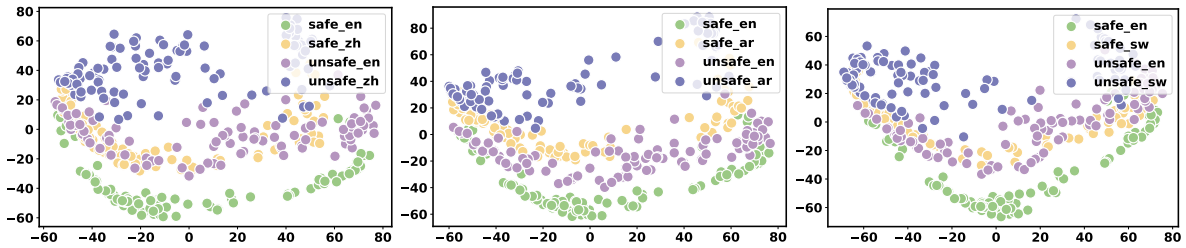
- MAPO constructs multilingual preference data by sampling multiple responses from an LLM in a given target languages and ranking them based on an alignment score computed via an external translation model, which measures their consistency with the response in the dominant language. The ranked responses form preference pairs that are then optimized using DPO (Rafailov et al., 2023).
- LIDR relies on the LLM's own translation capability to convert English preference data into target languages, followed by DPO optimization with an additional NLL loss.

While both methods explore multilingual preference data generation, they do not propose improvements to the multilingual preference optimization process itself, instead relying solely on DPO.

To evaluate whether MPO remains effective when trained on preference data obtained using these methods, we conduct experiments using MAPO- and LIDR-generated data. As shown in Table 16, MPO consistently achieves the best multilingual safety alignment results across both data sources, demonstrating its robustness to variations in preference data. These results emphasize that while MAPO and LIDR explore multilingual preference data construction, they do not address the fundamental challenges of multilingual preference optimization. MPO, in contrast, not only adapts to different multilingual data sources but also improves the optimization process, ensuring stable and effective multilingual safety alignment.

H.5 Visualization Analysis

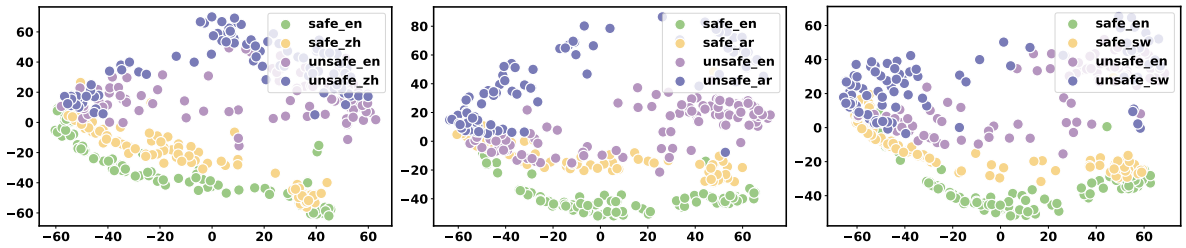
To further understand what MPO brings for the multilingual safety alignment of LLMs, as shown in Figure 6, we perform Principal Component Analysis (PCA) to visualize the multilingual representations in the activation spaces. Specifically, the multilingual harmful inputs are sourced from the AdvBench-X dataset. For each input, we append both a corresponding safe response and an unsafe response to visualize the model's representation. All representations are extracted from the final layer of the model's output and the backbone model



(a) The visualization of multilingual representations for English and Chinese on LLaMA-3.1-8B-Instruct.

(b) The visualization of multilingual representations for English and Arabic on LLaMA-3.1-8B-Instruct.

(c) The visualization of multilingual representations for English and Swahili on LLaMA-3.1-8B-Instruct.



(d) The visualization of multilingual representations for English and Chinese on LLaMA-3.1-8B-Instruct after multilingual safety alignment via MPO.

(e) The visualization of multilingual representations for English and Arabic on LLaMA-3.1-8B-Instruct after multilingual safety alignment via MPO.

(f) The visualization of multilingual representations for English and Swahili on LLaMA-3.1-8B-Instruct after multilingual safety alignment via MPO.

Figure 6: The visualization of multilingual representations for safe and unsafe inputs across different languages. The upper row (a-c) illustrates the representation space of the original LLaMA-3.1-8B-Instruct model for English (En) and three additional languages: Chinese (Zh), Arabic (Ar), and Swahili (Sw). The lower row (d-f) presents the corresponding representation space after applying multilingual safety alignment via MPO. The visualizations highlight the structural changes in the representation space induced by MPO alignment.

	X-AdvBench							
	En	**Zh**	**Jp**	**Ko**	**Ar**	**Bn**	**Sw**	**AVG.**
Aya-101	17.88	20.00	42.69	45.38	20.32	58.08	53.93	36.90
+ MPO	**8.65**	**14.04**	**20.19**	**25.19**	**9.81**	**31.35**	**39.23**	**21.21**

Table 17: Results of the multilingual safety performance on X-AdvBench. The backbone is Aya-101. The best results are highlighted in bold.

is LLaMA-3.1-8B-Instruct.

Notably, in all cases, the boundary between safe and unsafe inputs in English remains consistently clear. However, in the original model (a-c), the distinction between safe and unsafe inputs in the target languages (Zh, Ar, Sw) appears less structured and more entangled. After applying MPO alignment (d-f), the model demonstrates a significantly improved separation of safe and unsafe inputs in the target languages. This indicates that MPO enhances the multilingual safety alignment of the model, allowing it to develop clearer decision boundaries in languages beyond English.

H.6 Results on Aya-101

We includes results on an explicitly multilingual model Aya-101, providing more valuable empirical insights. The results are summarized in Table 17.

Although Aya-101 supports a wide range of languages, its multilingual safety alignment is still limited. After applying our MPO method—with English as the dominant language, given its well-established safety alignment—we observe a significant improvement in safety performance on X-Advbench across all 7 languages.