Language Chain in Alignment: Cross-Lingual Ranking Preference Optimization

Seungyoon Lee, Minhyuk Kim, Jungseob Lee, Heuiseok Lim*
Department of Computer Science and Engineering, Korea University
{dltmddbs100, mhkim0929, omanma1928, limhseok}@korea.ac.kr

Abstract

The alignment of Large Language Models heavily relies on English-centric high-quality preference data, which often leads to suboptimal performance in other languages. In this paper, we propose Cross-Lingual Ranking Preference Optimization (CRPO), a novel framework that leverages robust preference knowledge from English to facilitate preference alignment in the target language. We design a hierarchical structure within parallel preference pairs across the target language and English to jointly optimize intra- and inter-lingual preferences, thereby enhancing language adaptation and output quality. Building on the LambdaLoss framework, CRPO goes beyond the binary comparison based optimization by providing a relative ranking signal across multiple candidate responses. Our experiments across five languages with varying resource scales demonstrate that CRPO consistently outperforms standard approaches in both instruction-following and knowledge utilization capability. Notably, the robust performance gains observed across various weighting schemes further validate the empirical effectiveness of our hierarchical design in a multilingual setup. Furthermore, our findings highlight that CRPO significantly improves both reward margins and the log-probability of desirable responses, contributing to a more stable preference manifold for cross-lingual alignment.

1 Introduction

Integrating human feedback through the alignment process has emerged as a critical paradigm for optimizing the instruction-following capabilities of Large Language Models (LLMs) and grounding their objective functions in human value systems (Stiennon et al., 2020; Askell et al., 2021; Hong et al., 2024; Meng et al., 2024). In particular, Direct Preference Optimization (DPO) (Rafailov et al., 2023) has streamlined the complex Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) pipeline, effectively guiding models toward preferred responses while improving helpfulness and utility (Kirk et al., 2023; Pang et al., 2024; Tu et al., 2025). However, these alignment gains remain largely confined to English-centric contexts, leaving consistency and response accuracy in other languages as a formidable challenge. This disparity often leads to undesirable behavior, including input-output language inconsistency and degraded performance for other language queries (Lai et al., 2023a; Puttaparthi et al., 2023; Huang et al., 2023; Wu et al., 2024a; Marchisio et al., 2024).

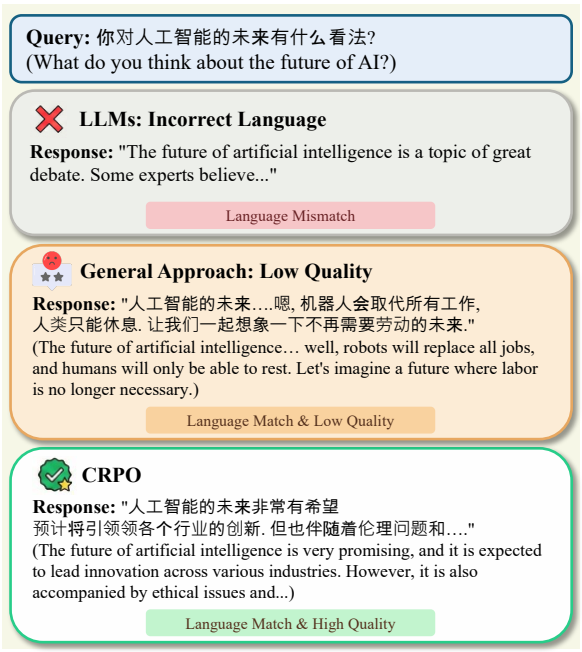


Figure 1: Example of model responses to multilingual prompt after alignment tuning.

Existing approaches to multilingual alignment typically rely on expensive human annotation or on integrating target language data into English-dominant corpora (Lai et al., 2023b; Zhao et al., 2024; Ranaldi et al., 2024; Lai et al., 2024). How-

*Corresponding author

ever, these approaches are limited by their inability to leverage the model's inherent English capabilities for cross-lingual alignment. By leaving this internalized English preference knowledge decoupled from the target language, they struggle to achieve effective cross-lingual alignment.

In this paper, we propose **C**ross-Lingual **R**anking **P**reference **O**ptimization (CRPO), a novel alignment framework that aligns the preference distribution of a target language by leveraging the model's inherent preference knowledge in English. CRPO moves beyond conventional binary comparisons within a single language; instead, we adopt an alignment strategy based on a hierarchical ranking structure that integrates two core dimensions: language consistency and response quality. As illustrated in Figure 1, while LLMs often suffer from language confusion (e.g., responding in English to a non-English prompt) and existing approaches generate low-quality outputs, CRPO guides the model toward generating high-quality responses appropriate for the target language query.

To achieve this, we transplant the "Learning-to-Rank" (LTR) (Liu et al., 2009) philosophy, which optimizes relative importance among multiple candidates, into the cross-lingual alignment training. To be specific, we design a hierarchical signal based on parallel preference pairs between English and the target language. This design facilitates alignment of target language preferences by referring to well-defined preference distributions, while inducing the generation of target language responses rather than English.

Our experiments show that CRPO achieves substantial improvements in instruction-following and knowledge-intensive tasks. Furthermore, our comparative analysis of reward and log-likelihood distributions demonstrates that CRPO leads to a more accurate likelihood preference of winning and losing responses on a held-out test set, which in turn translates to a better policy model for the target language. Our contributions are as follows:

- We propose CRPO, a novel framework that optimizes the hierarchical priority between language and quality by integrating ranking principles into cross-lingual preference alignment.
- Through an analysis of reward and probability distributions, we find that CRPO more effectively steers the preference distribution of the target language toward a desirable alignment compared to existing methodologies.
- We demonstrate that our cross-lingual preference hierarchy structure itself, which leverages the relationship between English and the target language, induces significant performance gains in target language alignment.

2 Related Work

Enhancing the multilingual capabilities of LLMs is a critical endeavor to ensure that global users can benefit from technological advancements without language barriers (Costa-Jussà et al., 2022; Zhao et al., 2024; Huang et al., 2024; Li et al., 2024a). Most research primarily focused on machine translating English datasets or training on heterogeneous mixtures of multilingual corpora (Workshop et al., 2022; Muennighoff et al., 2023; Lai et al., 2023b; Gao et al., 2024; Kew et al., 2024). However, these approaches often require large-scale datasets or suffer from persistent language bias (Aggarwal et al., 2024; Li et al., 2024c; Lee et al., 2025).

Recent studies seeking to mitigate these limitations have increasingly explored extending alignment algorithms proven in English to multilingual contexts (Yang et al., 2024; Dang et al., 2024a). In prior work, Wu et al. (2024b) facilitates zero-shot cross-lingual transfer in reward model training, aiming to build robust multilingual reward systems. MAPO (She et al., 2024) targets improvement in the multilingual reasoning domain using reward signals derived from external translation models. Similarly, MPO (Zhao et al., 2025) directly optimizes the reward margin gap across languages to enhance safety. While effective, these approaches rely on external dependencies, such as online machine translation pipelines, and are limited to specific domains. Another line of work operates directly on the policy model using implicit binary comparisons (Yang et al., 2025; Lee et al., 2025), yet they remain constrained by binary optimization, failing to capture complex preference distributions.

Although some works have explored list-based ranking to move beyond binary comparisons, these studies primarily focus on improving quality or safety and are strictly limited to English-centric contexts (Yuan et al., 2023; Song et al., 2024; Liu et al., 2025b). In contrast, CRPO distinguishes itself by optimizing a cross-lingual hierarchical ranking structure that jointly models language and quality. By enabling simultaneous modeling of linguis-

tic and quality hierarchies, CRPO leads to meaningful improvements in target language capabilities.

3 CRPO Framework

Our approach is established on the DPO framework, extending its pairwise alignment capabilities into a ranking paradigm specifically designed for cross-lingual contexts.

3.1 Preliminary: Direct Preference Optimization

DPO (Rafailov et al., 2023) is a stable and efficient method to align LLMs with human preferences without the need for training an explicit reward model or using complex reinforcement learning pipelines such as PPO (Schulman et al., 2017):

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right] \quad (1)$$

where π_θ is the policy model, π_{ref} is the reference policy given (x, y_w, y_l) are preference pairs which are the prompt, chosen response, and rejected response from the preference dataset. By minimizing the given loss, the model increases the margin between the preferred response y_w relative to y_l.

3.2 LambdaLoss Ranking Objective

Given that our aim is to encourage cross-lingual adaptation alongside robust preference alignment, the binary formulation of DPO is inherently limited, as it cannot accommodate multiple optimization spheres. To address this, we draw inspiration from LambdaLoss (Wang et al., 2018), a type of Learning to Rank (LTR) (Liu et al., 2009) framework that enables joint optimization of multiple items across a global ranking structure.

This approach allows the model to optimize well-founded ranking metrics, such as Discounted Cumulative Gain (DCG) (Burges et al., 2006), by weighting pairwise transitions based on their impact on the overall list. Specifically, for a prompt x and a set of candidate responses $\mathbf{y} = \{y_1, \ldots, y_n\}$, the training objective of our ranking scheme is defined as:

$$\mathcal{L}_{\text{lambda}}(\pi_\theta; \pi_{\text{ref}}, \beta) = \mathbb{E}_{(x,\mathbf{y},\psi)\sim\mathcal{D}} \left[\sum_{\psi_i>\psi_j} \Delta_{i,j} \log(1 + e^{-(s_i - s_j)})\right] \quad (2)$$

where $s_i = \beta \log \frac{\pi_\theta(y_i|x)}{\pi_{\text{ref}}(y_i|x)}$ represents the implicit reward score of response y_i. The term $\Delta_{i,j}$ is the Lambda weight, which determines the importance of the pair (y_i, y_j) based on their positions in the ranking:

$$\Delta_{i,j} = |G_i - G_j| \cdot \left|\frac{1}{D(\tau(i))} - \frac{1}{D(\tau(j))}\right| \quad (3)$$

In this formulation, G is the gain function (typically $G_i = 2^{\psi_i} - 1$ where ψ_i is the relevance label), and D is the rank discount function (typically $D(\tau(i)) = \log(1+\tau(i))$). Here, $\tau(i)$ denotes the rank position of y_i induced by the score s.

This objective can optimize the DCG metric by prioritizing the placement of high-relevance items at the top of the list by applying a logarithmic discount to lower ranks. Consequently, any rank inversion that displaces a desirable response from its superior position results in a significant reduction of the overall score, which allows the $\Delta_{i,j}$ to impose a stringent penalty on misalignment that disrupts the intended preference order.

3.3 CRPO: Cross-Lingual Ranking Preference Optimization

Building upon the LambdaLoss framework, CRPO aims to induce both cross-lingual and intra-lingual alignment by jointly optimizing target language responses alongside semantically equivalent preference knowledge from English. For each training instance, we construct a quadruple of responses consisting of: y_w^t, y_l^t, y_w^e, and y_l^e, which are chosen and rejected responses in the target language and English, respectively. To ensure a consistent preference manifold is maintained across languages, we assign hierarchical relevance labels (ψ) based on both response quality and language type among these four responses:

$$\psi(y_w^t) > \psi(y_w^e) > \psi(y_l^t) > \psi(y_l^e) \quad (4)$$

Under this hierarchy, the model is trained to prioritize the preferred response in the target language, y_w^t, as the highest-ranked candidate. At the same time, CRPO provides comprehensive supervisory signals by aligning all pairs within the $\{y^t, y^e\}$ set for a given target language prompt x. The model is compelled to jointly optimize both intra-lingual preferences ((y_w^t, y_l^t), (y_w^e, y_l^e)) and cross-lingual preference pairs, such as (y_w^t, y_l^e) and (y_w^e, y_l^t).

Since the optimization of target language involves concurrent alignment with semantically

equivalent English pairs, the model can leverage these English candidates as informative cues. Furthermore, prioritizing y_w^t provides training signals for maintaining input-output language consistency. Thus, the gain function with ψ in CRPO enables the model to inherit the logical consistency of English preferences while imposing a stringent quality constraint on responses within the target language. The gain functions for each response in CRPO are detailed in Appendix A.

Finally, we employ a negative log-likelihood (NLL) loss term for y_w^t in our loss to improve alignment performance during the instruction tuning process. The resulting loss function of CRPO is defined as:

$$\mathcal{L}_{\text{CRPO}} = (1-\alpha) \cdot \mathcal{L}_{\text{NLL}} + \alpha \cdot \mathcal{L}_{\text{lambda}}(\pi_\theta; \pi_{\text{ref}}, \beta) \tag{5}$$

where α is a hyperparameter that balances the ranking-based alignment and the language modeling objective. While conventional DPO is restricted to binary comparisons of responses within the target language, CRPO promotes the precise mapping of hierarchies defined by both language and quality onto a unified optimization plane.

3.4 Weighting Schemes for CRPO

The flexibility of the LambdaLoss framework allows CRPO to incorporate various weighting functions for each optimizing different bounds of the ranking metric. We explore three schemes for $\Delta_{i,j}$ derived from LambdaLoss to investigate the effectiveness in cross-lingual alignment.

LambdaRank optimizes a coarse upper bound of the nDCG metric. It defines the weight based on the absolute difference between the inverse discounts of the two positions, which is presented in Equation 3. This scheme provides a foundational gradient signal for ranking by considering the global positions of the responses in the list.

nDCG2 introduces a tighter bound for nDCG optimization. Unlike the global position-based weighting of LambdaRank, it focuses on the distance between the two items being compared:

$$\Delta_{i,j}^{\text{nDCG2}} = \left|G_i - G_j\right| \cdot \left| \frac{1}{D(|\tau(i) - \tau(j)|)} - \frac{1}{D(|\tau(i) - \tau(j)| + 1)} \right| \tag{6}$$

where $|\tau(i) - \tau(j)|$ represents the rank distance between the two responses. This scheme has been empirically shown to achieve superior performance by providing more localized and precise gradients for rank inversion.

nDCG2++ is a hybrid scheme that combines the strengths of both nDCG2 and LambdaRank to maximize ranking performance:

$$\Delta_{i,j}^{\text{nDCG2++}} = \mu \cdot \Delta_{i,j}^{\text{nDCG2}} + \Delta_{i,j} \tag{7}$$

where μ is a hyperparameter that controls the contribution of the nDCG2 component. In our experiments, μ is set to 10 following the prior study (Wang et al., 2018).

While nDCG2++ is known to reach the strongest performance in traditional Information Retrieval tasks, we found that nDCG2 or LambdaRank alone can be sufficient for aligning LLM preference distributions across languages. To demonstrate the generality and theoretical robustness of the CRPO framework, we employ nDCG2 as our primary weighting scheme for the experiments.

4 Experimental Setup

4.1 Models and Training

Models We perform preference optimization with three different models, Llama-2-7B (Touvron et al., 2023), Llama-3-8B (Dubey et al., 2024), and Mistral-7B-v0.1 (Jiang et al., 2023). We employ five languages with varying resource availability: Chinese (zh), Indonesian (id), Korean (ko), Swahili (sw), and Bengali (bn). They were chosen to provide a mix of high-, medium-, and low-resource languages, typological and script diversity, while satisfying the practical constraints of available evaluation datasets.

Dataset To construct the preference dataset for each language, we sample 3,000 instances from the UltraFeedback dataset (Cui et al., 2024) and translate them into the respective target languages using `gpt-5-chat`. This procedure yields fully parallel target language pairs (x_t, y_w^t, y_l^t) that correspond directly to the original English preference pairs (x_e, y_w^e, y_l^e). We randomly split the resulting dataset, including English pairs, into a 90% training set and a 10% test set.

Baseline We compare CRPO against two competitive baselines that aim to align cross-lingual preferences via monolithic strategy. We establish SFT+DPO as our strong baseline. In this setup, the model is optimized using response pairs in the language of the input. Since the training dataset

encompasses both English and the target language, the model is trained on (y_w^t, y_l^t) for x^t and (y_w^e, y_l^e) for x^e, in the same batch.

As another method, CLO (Lee et al., 2025) is a variant of DPO that incorporates binary comparison between English and target language response pairs within the same batch. To be specific, CLO explicitly steers the model toward generating desirable and language-consistent responses by designating the target language response as chosen and the English response as rejected. Furthermore, it mitigates inherent English bias and facilitates alignment by restricting the NLL loss exclusively to target language responses.

Note that all baselines and models use identical hyperparameters and the same amount of parallel datasets to ensure a fair comparison. Further details about the training can be found in Appendix B.

4.2 Evaluation

AlpacaEval We employ AlpacaEval (Li et al., 2023) to evaluate models' conversational ability across a diverse set of queries. AlpacaEval comprises 805 questions from 5 datasets and evaluates model performance through qualitative comparisons of reference responses and candidate model outputs. To support evaluation across multiple languages, we adopt the multilingual version of AlpacaEval established in prior work (Lee et al., 2025) as our primary evaluation suite. We report the win rate (WR) and length-controlled win rate (LC) against the SFT model. The LC metric is specifically designed to be robust against model verbosity (Dubois et al., 2024). Also, by including the instruction in the evaluation prompt, the model is penalized for responding in a language other than the target language. We use `gpt-5-chat` as a judge for assessment and also report multilingual Arena-Hard results in Appendix C.

Knowledge & Understanding We further evaluate the models' capacity for knowledge utilization and linguistic comprehension. We use MMMLU (Multilingual Massive Multitask Language Understanding) (Hendrycks et al., 2021) to verify whether the trained model can leverage inherent world knowledge in target language contexts. MMMLU is a human-translated multilingual extension of the original MMLU benchmark, encompassing 57 subjects across STEM, humanities, and social science [1]. In addition, we employ Belebele (Bandarkar et al., 2024), a benchmark for multilingual reading comprehension. In this task, we can evaluate literacy and contextual understanding, as the model should select the most appropriate answer from four choices based on the short passage and question. Following standard practice, we report accuracy by treating the option with the highest overall log-likelihood as the model's predicted answer (Gao et al., 2021).

In-depth Analysis To move beyond aggregate benchmark scores and gain deeper insights into the model's internal mechanics, we conduct an intrinsic analysis. We investigate the reward distributions and log probabilities of the preference pairs in our test set to verify whether each method enables balanced alignment that reinforces desirable generative behaviors, rather than relying solely on suppressing suboptimal outputs.

Furthermore, we measure the quality distribution of generated responses using an external reward model. The objective of this analysis is to empirically demonstrate that the internal optimization of CRPO translates into measurable improvements in output quality, ensuring that the model's preference manifold is robustly aligned with objective judgments in the target language.

5 Results and Analysis

In this section, we present the main results of our experiments, highlighting CRPO's superior performance across various benchmarks in Section 5.1. Then in Section 5.2, we provide an in-depth understanding of the internal reward dynamics and generative behaviors that distinguish CRPO from other alignment methods. In addition, we employ an external reward model to objectively quantify improvements in absolute generation quality in Section 5.3. Finally, we conduct a study on various ranking schemes to investigate how different configurations of the hierarchical preference structure influence the effectiveness of cross-lingual alignment in Section 5.4.

5.1 Main Results

Overall Performance As shown in Table 1, all preference optimization methods provide improvements over the SFT model in most cases, with CRPO achieving the best results across languages. SFT+DPO, a standard approach for model alignment, also serves as a robust baseline in other languages. CLO exhibits comparable or slightly below

[1] https://huggingface.co/datasets/openai/MMMLU

Model	Method	Chinese		Indonesian		Korean		Swahili		Bengali	
		LC (%)	WR (%)	LC (%)	WR (%)	LC (%)	WR (%)	LC (%)	WR (%)	LC (%)	WR (%)
Llama-2-7B	SFT+DPO	57.80	58.50	57.92	58.88	54.28	55.40	38.63	38.07	34.90	35.68
	CLO	56.41	56.08	53.84	53.91	51.15	50.24	**44.49**	**43.97**	33.15	30.93
	CRPO	**64.18**	**64.34**	**63.83**	**64.47**	**63.02**	**63.97**	44.07	43.72	**37.52**	**36.77**
Llama-3-8B	SFT+DPO	60.24	60.37	59.24	60.37	64.97	65.96	46.23	46.95	34.77	36.89
	CLO	58.70	58.13	52.42	52.96	53.94	53.41	43.78	44.93	35.67	33.29
	CRPO	**62.45**	**62.36**	**67.97**	**68.40**	**68.45**	**68.81**	**61.84**	**62.17**	**53.86**	**55.65**
Mistral-7B	SFT+DPO	55.65	56.14	55.98	58.50	51.49	52.85	27.95	27.33	24.63	24.59
	CLO	55.18	55.03	51.29	51.98	48.14	48.32	35.15	35.40	19.98	18.01
	CRPO	**62.83**	**63.10**	**66.81**	**68.19**	**62.90**	**63.72**	**50.76**	**51.98**	**48.69**	**48.57**

Table 1: AlpacaEval across five target languages. LC and WR denote length-controlled and raw win rate, respectively. The best results for each model are highlighted in **bold**.

Model	Lang	SFT+DPO	CLO	CRPO
MMMLU				
Llama-2-7B	id	26.33	24.69	**27.85**
	ko	26.20	28.57	**31.07**
	sw	25.08	**25.62**	25.12
Llama-3-8B	id	41.69	36.41	**45.94**
	ko	39.58	**40.19**	39.84
	sw	36.02	27.48	**37.03**
Mistral-7B	id	37.36	31.38	**37.47**
	ko	30.39	29.27	**35.01**
	sw	30.57	28.32	**31.78**
Belebele				
Llama-2-7B	id	33.11	36.55	**36.66**
	ko	33.00	35.66	**36.33**
	sw	29.22	28.22	**29.22**
Llama-3-8B	id	64.22	35.11	**65.88**
	ko	57.55	59.44	**68.66**
	sw	47.55	28.66	**49.22**
Mistral-7B	id	**48.33**	30.44	47.44
	ko	49.66	35.77	**51.77**
	sw	33.00	33.88	**34.77**

Table 2: MMMLU and Belebele evaluation results. We report the performance of zero-shot in MMMLU and one-shot in Belebele.

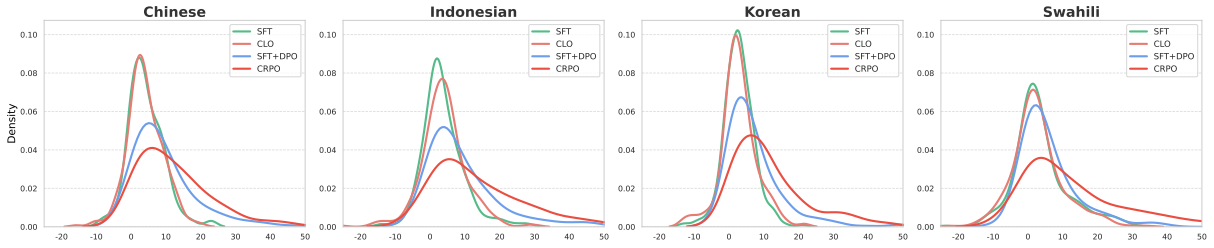
performance to SFT+DPO, demonstrating the utility of a straightforward objective that suppresses the probability of English responses. Furthermore, we include the English performance of each case in Appendix D in terms of alignment tax.

Robustness in Low-Resource Languages While most methods outperform SFT in high- and mid-resource languages, a different tendency emerges in low-resource settings. Unlike CRPO, other methods exhibit notable performance degradation in low-resource scenarios. This trend is particularly pronounced for Llama-3-8B, which suffers from a collapse compared to the SFT, as evidenced by notably low WR in Bengali.

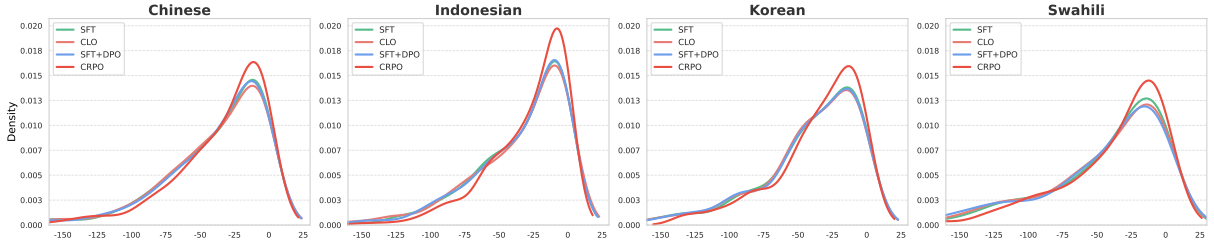
This stems from optimization instability and excessive parameter volatility during the preference alignment when the model's linguistic grounding in the target language is insufficient (Kadyrbek et al., 2025; Lee et al., 2025). From this perspective, SFT+DPO, relying on isolated preference pairs in the target language, fails to establish a stable preference manifold in low-resource scenarios. Similarly, the performance drop in CLO shows that simply enforcing language-consistent responses through binary cross-lingual hierarchies is insufficient for robust alignment.

While CRPO also encounters adaptation challenges due to the inherent limitations of models, it consistently surpasses baselines and demonstrates remarkable robustness in most cases. Notably, CRPO achieves a significant performance gain in Swahili with Llama-3, scoring a WR of 62.17. This gain can be attributed to the use of relatively well-aligned English pairs as a logical anchor for aligning the target language. By realigning language quality and hierarchy, CRPO exhibits the most robust resistance to this alignment collapse and contributes to overall alignment stabilization.

Knowledge Utilization The validity of CRPO extends beyond conversational improvements, as evidenced in the knowledge utilization and linguistic comprehension areas. As shown in Table 2, CRPO achieves competitive performance across both benchmarks. Notably, the MMMLU score in Indonesian for Llama-3 exceeds the strongest baseline by over 4 points, while the Korean Belebele score reaches 68.66. These results demonstrate



(a) Reward Difference



(b) Log Probability Distribution

Figure 2: (a) Reward difference distribution and (b) Log-likelihood distribution on chosen responses (y_w) of Llama-3-8B trained with SFT (blue), CLO (orange), SFT+DPO (blue) and CRPO (red) across languages.

Model	Method	zh	id	ko	sw
Llama-2-7B	SFT	-1.749	-1.797	-2.870	-2.901
	SFT+DPO	-1.227	**-0.904**	-2.121	-2.958
	CLO	-1.539	-1.200	-2.855	-2.457
	CRPO	**-0.607**	-0.936	**-1.601**	**-2.175**
Llama-3-8B	SFT	0.344	0.208	-0.904	-0.487
	SFT+DPO	1.506	1.585	0.869	-0.212
	CLO	0.613	1.034	-0.119	-0.531
	CRPO	**1.738**	**1.883**	**0.911**	**1.122**
Mistral-7B	SFT	-0.706	-0.808	-1.800	-0.667
	SFT+DPO	0.255	-0.037	-0.719	-1.519
	CLO	-0.096	-0.665	-1.237	-0.981
	CRPO	**0.508**	**0.805**	**-0.344**	**-0.040**

Table 3: Average of reward scores from Skywork-Reward-V2-Qwen3-8B on generation outputs. The best results achieved across the methods are highlighted in bold, while the second-best results are underlined.

that the ranking structure in CRPO successfully anchors the model's inherent knowledge system within the target language space, thereby achieving better alignment for complex reasoning tasks.

5.2 Reward and Probability Shifts

To investigate the impact of CRPO on the model's preference manifold, we examine the internal dynamics of reward differences and generative log-probabilities. As illustrated in Figure 2, the magnitude of distribution shifts differs across methods regarding the SFT model as a default.

In Figure 2 (a), we find that regardless of the language, CRPO consistently achieves a large reward difference and shifts the distribution in a positive direction compared to the SFT. Likewise, the SFT+DPO approach exhibits a similar trend, though the magnitude of the shift is relatively smaller. In contrast, CLO yields almost identical reward difference to SFT, or in the case of Indonesian, even shows signs of degradation.

One of the primary risks in preference optimization is the improper focus on maximizing the reward margin, which often results in suppressing the probability of rejected responses rather than increasing the probability of chosen responses. This imbalance can undermine the language model's inherent stability and fluency by failing to provide an explicit guide for target responses, potentially leading to degeneration (Meng et al., 2024; Hong et al., 2024). As shown in the comparison of Figure 2 (a) and (b), while other methods increase the reward margin, their log-likelihood distributions for chosen responses remain largely congruent with the SFT. This indicates that other approaches primarily fail to incentivize the generation of favored responses.

On the other hand, CRPO assigns higher values to both the log-likelihood of the chosen response and the reward difference across all languages than in other cases. This demonstrates that CRPO achieves better alignment in the target language by increasing the likelihood of the favored response. Further analysis of the reward distribution is provided in Appendix E.

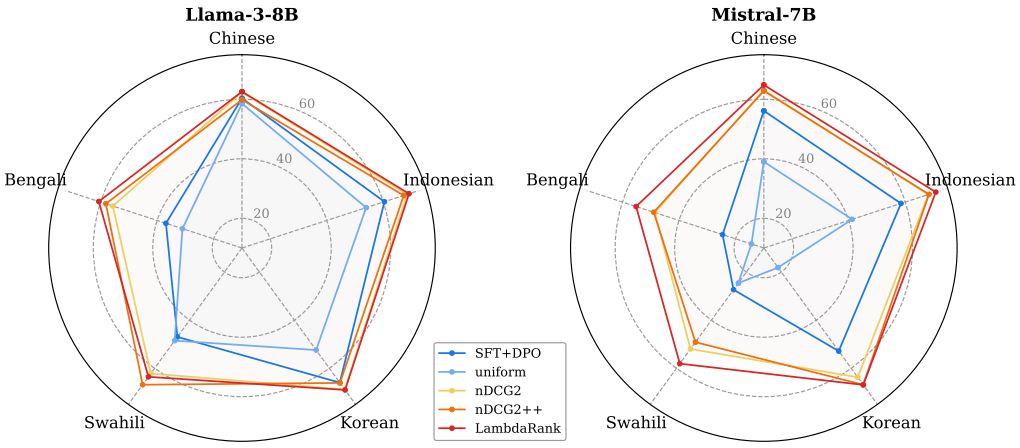


Figure 3: Win rate comparison of different weighting schemes within the CRPO framework.

5.3 External Quality Assessment

While win rates provide a measure of relative preference, they may not fully capture the absolute quality shift in the model's generative distribution. To provide a more objective assessment of response quality, we use Skywork-Reward-V2-Qwen3-8B (Liu et al., 2025a), an external reward model known to support multilingual environments, to score the outputs of each method on the test set. Table 3 summarizes the average reward scores across diverse models and languages.

The results demonstrate that CRPO achieves superior reward magnitudes across nearly all configurations. A notable score escalation is observed in the Mistral-7B; in Indonesian (id), while the SFT+DPO reaches a score of -0.037, CRPO elevates the quality to a considerable positive margin of 0.805, outperforming SFT+DPO by a wide gap. This also substantiates our claim that anchoring target language alignment to established English preference structures can guide the model toward higher-quality regions of the generation manifold.

5.4 Study on Ranking Scheme

We investigate how the choice of weighting scheme in Equation 2 influences cross-lingual preference alignment. Figure 3 shows AlpacaEval results obtained by diverse weighting schemes within the same CRPO framework.

As shown in Figure 3, all three schemes consistently outperform both SFT+DPO and the uniform scheme, confirming that hierarchical ranking signals provide more informative supervision than pairwise comparisons. We can attribute the success across diverse configurations to CRPO's hierarchical structural design. As evidence for this, assigning a uniform weight to all response pairs results in a notable performance drop. The degradation mainly arises from the lack of alignment between the languages. Removing hierarchical cues triggers an English bias, leading to a failure in linguistic consistency with the given prompt.

Moreover, LambdaRank and nDCG2 achieve strong performance across most settings. This suggests that global position-based signals and local distance-based signals can complement each other depending on the model and language. Given that our experiments are conducted under the nDCG2 scheme, the additional gains observed with alternative schemes further indicate that the potential performance ceiling of CRPO may be even higher.

6 Conclusion

In this paper, we propose CRPO, a new alignment framework designed to establish robust preference manifolds both within and across languages. By reformulating the target language adaptation and alignment task as a hierarchical ranking problem, CRPO benefits from the transfer of preference knowledge along the language chain. Our empirical results demonstrate that CRPO not only achieves consistent superiority over standard pairwise approaches but also enhances the generative likelihood of preferred responses, facilitating the generation of high-fidelity, precise responses in the target language. We believe that our approach shifts the paradigm of cross-lingual alignment from isolated binary comparisons to an integrated cross-lingual ranking perspective. Moving forward, we will investigate a dynamic gain adaptation strategy based on linguistic similarities and training stages. By mitigating the optimization complexity inherent in

expansive ranking spaces, we expect to ensure robust cross-lingual alignment across an even wider array of language pairs.

Limitations

Although we use the same amount of the dataset during the training, our proposed framework employs a hierarchical ranking structure that processes four response candidates in a single step. This inherently demands more computational resources during training. However, this overhead can be considered an unavoidable cost inherent to methods that leverage item-wise ranking, not only limited to the preference optimization field. Given the limitations of standard binary comparisons in optimizing both quality and consistency simultaneously, the performance gains achieved by CRPO demonstrate that this additional cost is a worthwhile trade-off for effective cross-lingual alignment.

Furthermore, our main experiments include evaluation results from the multilingual version of the AlpacaEval dataset, translated using gpt-4o in prior work (Lee et al., 2025), and the MMMLU dataset (Hendrycks et al., 2021), which was translated into the respective languages by professional human translators. Additionally, we employed the Belebele dataset (Bandarkar et al., 2024), which was constructed and reviewed by human experts during translation. While these datasets are suitable for measuring general model performance, they may not fully capture the model's ability to respond appropriately to queries involving language-specific characteristics or cultural contexts. Although this limitation may prevent a comprehensive evaluation of the model's grasp of cultural nuances and localized content, we made every effort to utilize the highest-quality benchmarks available.

Ethics Statement

In this research, we utilized only publicly available resources. All training and evaluation data were obtained from publicly authorized open-access repositories. Nevertheless, we recognize the potential for residual biases or harmful concepts in the English source data to be propagated across languages. To mitigate this, we carefully adhered to the copyright guidelines and terms of use for all original works, datasets, and language resources, including translated materials. We confirm that the collection, processing, and application of these resources raise no distinct ethical concerns.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2021R1A6A1A03045425). This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI). This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2026 (Project Name: Development of an AI Agent Integrating Korean Language Knowledge for Personalized Language Consultation Services, Project Number: RS-2026-25506607, Contribution Rate: 25%). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the Leading Generative AI Human Resources Development (IITP-2026-RS-2026-25545512) grant funded by the Korea government (MSIT).

References

Divyanshu Aggarwal, Ashutosh Sathe, Ishaan Watts, and Sunayana Sitaram. 2024. MAPLE: Multilingual evaluation of parameter efficient finetuning of large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14824–14867, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Christopher Burges, Robert Ragno, and Quoc Le. 2006. Learning to rank with nonsmooth cost functions. *Advances in neural information processing systems*, 19.

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*.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, et al. 2024. Ultrafeedback: Boosting language models with scaled ai feedback. In *International Conference on Machine Learning*, pages 9722–9744. PMLR.

John Dang, Arash Ahmadian, Kelly Marchisio, Julia Kreutzer, Ahmet Üstün, and Sara Hooker. 2024a. RLHF can speak many languages: Unlocking multilingual preference optimization for LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13134–13156, Miami, Florida, USA. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024b. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

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*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Changjiang Gao, Hongda Hu, Peng Hu, Jiajun Chen, Jixing Li, and Shujian Huang. 2024. Multilingual pretraining and instruction tuning improve cross-lingual knowledge alignment, but only shallowly. 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 6101–6117, Mexico City, Mexico. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, et al. 2021. A framework for few-shot language model evaluation. *Zenodo*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

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

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 *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Kaiyu Huang, Fengran Mo, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, et al. 2024. A survey on large language models with multilingualism: Recent advances and new frontiers. *arXiv preprint arXiv:2405.10936*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Nurgali Kadyrbek, Zhanseit Tuimebayev, Madina Mansurova, and Vítor Viegas. 2025. The development of small-scale language models for low-resource languages, with a focus on kazakh and direct preference optimization. *Big Data and Cognitive Computing*, 9(5):137.

Tannon Kew, Florian Schottmann, and Rico Sennrich. 2024. Turning English-centric LLMs into polyglots: How much multilinguality is needed? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13097–13124, Miami, Florida, USA. Association for Computational Linguistics.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2023. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023a. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023b. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327.

Wen Lai, Mohsen Mesgar, and Alexander Fraser. 2024. LLMs beyond English: Scaling the multilingual capability of LLMs with cross-lingual feedback. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8186–8213, Bangkok, Thailand. Association for Computational Linguistics.

Jungseob Lee, Seongtae Hong, Hyeonseok Moon, and Heuiseok Lim. 2025. Cross-lingual optimization for language transfer in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15100–15119, Vienna, Austria. Association for Computational Linguistics.

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 8058–8076.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024b. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*.

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

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ninghao Liu, and Mengnan Du. 2024c. Quantifying multilingual performance of large language models across languages. *arXiv preprint arXiv:2404.11553*.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, et al. 2025a. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. *arXiv preprint arXiv:2507.01352*.

Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, Peter J Liu, and Xuanhui Wang. 2025b. LiPO: Listwise preference optimization through learning-to-rank. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2404–2420, Albuquerque, New Mexico. Association for Computational Linguistics.

Tie-Yan Liu et al. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331.

Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, and Sebastian Ruder. 2024. Understanding and mitigating language confusion in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6653–6677, Miami, Florida, USA. Association for Computational Linguistics.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

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.

Richard Yuanzhe Pang, Weizhe Yuan, He He, Kyunghyun Cho, Sainbayar Sukhbaatar, and Jason Weston. 2024. Iterative reasoning preference optimization. *Advances in Neural Information Processing Systems*, 37:116617–116637.

Poorna Chander Reddy Puttaparthi, Soham Sanjay Deo, Hakan Gul, Yiming Tang, Weiyi Shang, and Zhe Yu. 2023. Comprehensive evaluation of chatgpt reliability through multilingual inquiries. *arXiv preprint arXiv:2312.10524*.

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, and Andre Freitas. 2024. Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7961–7973, Bangkok, Thailand. Association for Computational Linguistics.

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

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, Bangkok, Thailand. Association for Computational Linguistics.

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

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021.

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

Songjun Tu, Jiahao Lin, Xiangyu Tian, Qichao Zhang, Linjing Li, Yuqian Fu, Nan Xu, Wei He, Xiangyuan Lan, Dongmei Jiang, et al. 2025. Enhancing llm reasoning with iterative dpo: A comprehensive empirical investigation. *arXiv preprint arXiv:2503.12854*.

Xuanhui Wang, Cheng Li, Nadav Golbandi, Michael Bendersky, and Marc Najork. 2018. The lambdaloss framework for ranking metric optimization. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 1313–1322.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Suhang Wu, Jialong Tang, Baosong Yang, Ante Wang, Kaidi Jia, Jiawei Yu, Junfeng Yao, and Jinsong Su. 2024a. Not all languages are equal: Insights into multilingual retrieval-augmented generation. *arXiv preprint arXiv:2410.21970*.

Zhaofeng Wu, Ananth Balashankar, Yoon Kim, Jacob Eisenstein, and Ahmad Beirami. 2024b. 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.

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. 2024. 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. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21125–21147, Vienna, Austria. Association for Computational Linguistics.

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*, 36:10935–10950.

Jun Zhao, Zhihao Zhang, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond english: An empirical study on language capability transfer. *arXiv preprint arXiv:2401.01055*.

Weixiang Zhao, Yulin Hu, Yang Deng, Tongtong Wu, Wenxuan Zhang, Jiahe Guo, An Zhang, Yanyan Zhao, Bing Qin, Tat-Seng Chua, and Ting Liu. 2025. MPO: Multilingual safety alignment via reward gap optimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23564–23587, Vienna, Austria. Association for Computational Linguistics.

A Gain Formulation in Cross-Lingual Ranking

Preference Hierarchy In our pilot study, we observed that language adaptation is more readily achieved than the fine-grained differentiation of response quality within the same language, which can easily overshadow quality optimization during training. To reflect this, we established the following preference hierarchy for our ranking-based objective: $\psi(y_w^t) > \psi(y_w^e) > \psi(y_l^t) > \psi(y_l^e)$. A critical design choice in this formulation is the relation $\psi(y_w^e) > \psi(y_l^t)$, which prioritizes a high-quality response in English over a lower-quality response in the target language. Furthermore, this provides a learning signal that encourages the model to generate desirable responses regardless of the language. Since gain adjustments across items in a ranking structure are interdependent rather than orthogonal, penalizing the English winner below the target language loser would otherwise heavily bias the learning signal toward language matching.

Balance between Two Dimensions This philosophy also led to our choice of gain values. In standard Information Retrieval, which adopts the LambdaLoss framework, the gain function typically fol-

Gain values (G)	Chinese		Korean		Indonesian		Swahili		Bengali	
	WC (%)	LC (%)	WC (%)	LC (%)	WC (%)	LC (%)	WC (%)	LC (%)	WC (%)	LC (%)
7, 3, 1, 0	60.50	60.75	66.46	66.04	67.20	66.86	66.46	66.35	54.16	52.54
9, 5, 7, 4	60.00	60.39	66.21	66.11	65.22	65.52	**67.14**	**67.43**	53.98	53.79
9, 7, 5, 4	**62.36**	**62.46**	**68.82**	**68.46**	**68.41**	**67.98**	62.17	61.85	**55.65**	**53.87**

Table 4: WR and LC according to different gain values.

Model	Method	Chinese		Indonesian		Korean		Swahili		Bengali	
		LC (%)	WR (%)	LC (%)	WR (%)	LC (%)	WR (%)	LC (%)	WR (%)	LC (%)	WR (%)
Llama-2-7B	SFT+DPO	55.06	55.71	55.89	56.52	57.79	58.26	56.04	56.46	51.46	52.42
	CLO	57.46	57.27	53.49	52.24	55.30	54.91	53.41	51.93	53.22	52.61
	CRPO	**62.33**	**62.36**	**60.41**	**60.62**	**60.46**	**60.87**	**60.43**	**60.00**	**61.60**	**61.99**
Llama-3-8B	SFT+DPO	**57.71**	**58.45**	60.31	61.30	56.23	56.83	58.06	58.01	56.98	**57.76**
	CLO	52.83	52.48	56.89	56.21	52.66	50.99	54.46	52.61	56.27	54.78
	CRPO	56.73	57.20	**64.24**	**64.35**	**59.19**	**59.07**	**60.53**	**60.06**	**57.43**	57.64
Mistral-7B	SFT+DPO	**61.30**	**65.59**	58.44	61.06	**63.31**	**69.69**	51.73	54.88	47.65	55.47
	CLO	51.84	54.47	52.05	51.61	58.19	61.80	50.37	49.81	47.97	45.83
	CRPO	50.60	57.08	**59.96**	**61.86**	57.42	64.35	**56.21**	**57.76**	**53.49**	**55.86**

Table 5: English performance in AlpacaEval benchmark. The best results for each model are highlighted in **bold**.

lows an exponential discount: $G_i = 2^{\psi_i} - 1$ (Wang et al., 2018). However, applying this to our dual-dimensional objective can introduce an inductive bias.

Given our hierarchy, the standard formula yields $G^t_w = 7$, $G^e_w = 3$, $G^t_l = 1$, and $G^e_l = 0$. Although ranking optimization inherently assigns disproportionate weights between items, this exponential spacing exacerbates biased local gradients.

To resolve this and provide a more balanced optimization manifold, we manually set the gains to $G^t_w = 9$, $G^e_w = 7$, $G^t_l = 5$, and $G^e_l = 4$. In this setup, the primary intra-language quality gaps ($G^t_w - G^t_l = 4$ for target, $G^e_w - G^e_l = 3$ for English) remain strictly greater than the primary language gap ($G^t_w - G^e_w = 2$). This constraint explicitly ensures that the LambdaLoss framework jointly optimizes both cross-lingual alignment and quality, preventing the simple language-matching signal from overwhelming the fine-grained quality optimization.

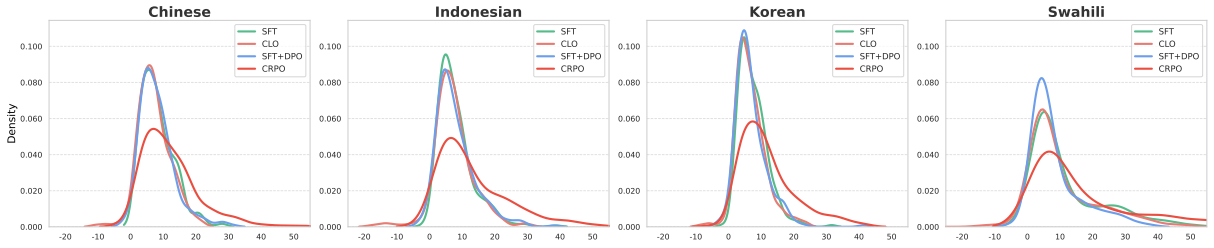
Empirical Validation To validate our gain formulation, we compare the performance of Llama-3-8B across different gain configurations, as presented in Table 4. The theoretical range of possible gains is infinite, but as it is computationally prohibitive to explore the entire space, our experiments are restricted to a set of representative fixed values.

The standard exponential setting $(7, 3, 1, 0)$ generally yields suboptimal win rates, suggesting that disproportionate local gradients hinder effective intra-lingual alignment. Furthermore, we examine a configuration $(9, 5, 7, 4)$ that explicitly violates our preference hierarchy by penalizing the English winner ($G = 5$) below the target language loser ($G = 7$). This configuration also yields slightly lower performance than our hierarchy across most languages, including Chinese, Korean, Indonesian, and Bengali, demonstrating that heavily biasing the signal toward language matching compromises overall response quality. Interestingly, a different tendency appears in Swahili. We attribute this to the low-resource nature of Swahili, where an aggressive language-matching penalty might provide a temporary empirical advantage in enforcing the target script. Nevertheless, our proposed hierarchy configuration demonstrates greater robustness across diverse linguistic environments, justifying our balanced hierarchical design.

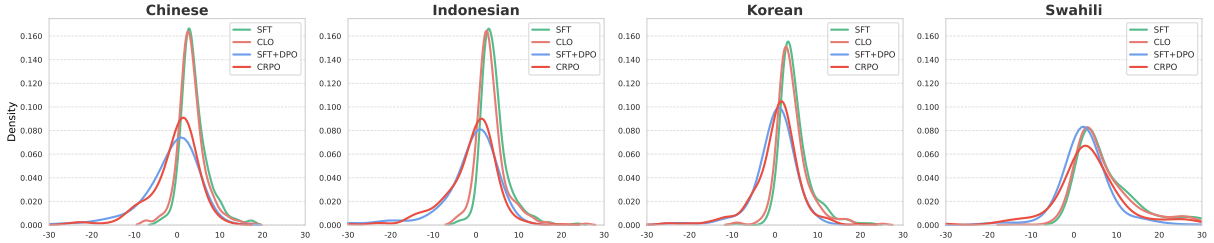
B Implementation Details

Training We observe that hyperparameter tuning is crucial for achieving optimal performance of preference optimization methods. To ensure a fair comparison, we conduct thorough hyperparameter tuning across all methods in our experiments.

We conduct preliminary experiments to search for learning rate in $[5e-7, 8e-6, 1e-5]$ and training epochs in $[1, 2]$. We find that an effective learning rate of 8e-6 and training for two epochs consistently yield stable results across all models



(a) Reward distribution of y_w



(b) Reward distribution of y_l

Figure 4: Reward distribution of chosen response (a) and rejected response (b) of Llama-3-8B trained with SFT (blue), CLO (orange), SFT+DPO (blue) and CRPO (red) across languages.

and languages. Therefore, we fix these values for all preference optimization experiments and report results using the final checkpoint. Additionally, we set the maximum sequence length to 3072 and apply a linear learning rate schedule with a 10% warmup ratio. To balance the training objective, we use a weighting factor of $\alpha = 0.2$ for CRPO and SFT+DPO, and 0.5 for CLO following the recommended value in the original paper. We set β to 0.1, the default value in DPO. All experiments are conducted under four NVIDIA A100 GPUs.

Evaluation For generative evaluation, we maintain the same maximum sequence length as used during the training phase to ensure consistency. In the generation process, we use temperature as the primary hyperparameter in the decoding step and set it to 0.8 to balance the trade-off between coherence and diversity in the outputs.

C Arena-Hard

To further substantiate our evaluation, we report the multilingual Arena-Hard (m-ArenaHard)[2] (Dang et al., 2024b) evaluation results in Table 6. The m-ArenaHard benchmark was created by translating the prompts from the original Arena-Hard-v0.1 (Li et al., 2024b) test dataset to 22 languages. We conduct supplementary experiments on the three languages that overlap with those covered in our study. Since several queries in the dataset exceed the relatively short context length of Llama-2 (4096 tokens), Llama-2 is excluded from the evaluation. Similar to AlpacaEval, we report the win rate against the SFT model.

[2] https://huggingface.co/datasets/CohereLabs/m-ArenaHard

Model	Method	Chinese	Indonesian	Korean
Llama-3-8B	SFT+DPO	61.56	61.34	64.30
	CLO	54.49	54.53	54.95
	CRPO	**62.15**	**61.87**	**68.81**
Mistral-7B	SFT+DPO	54.62	54.34	55.39
	CLO	55.49	53.34	53.02
	CRPO	**58.22**	**63.85**	**62.63**

Table 6: Win rate in Arena-Hard benchmark.

D Impact on Alignment Tax

In our main experiments, we focused on evaluating the target languages to demonstrate the effectiveness of cross-lingual alignment. However, since our approach leverages the model's inherent English knowledge as an anchor to induce transfer to the target language, it is imperative to investigate how the alignment training on the target language affects the model's English proficiency. To this end, we conduct an additional English evaluation on AlpacaEval to measure the alignment tax.

Table 5 presents the English performance of the models trained under each target language setting. Note that all methods also include English in their training data composition. As shown in the results, CRPO achieves higher English performance compared to the baselines in most cases, indicating that

it successfully mitigates the alignment tax. To be specific, for the Llama-2-7B and Llama-3-8B models, CRPO achieves the highest WR across almost all target language configurations. This suggests that CRPO preserves the robustness of the model's English capabilities while effectively performing cross-lingual alignment.

E Reward Distribution Analysis

In addition to Figure 2, we also provide an analysis of the reward distribution of responses in Figure 4. As shown in Figure 4 (a), CRPO consistently yields higher rewards for favored responses compared to the other methods, whereas the reward distributions for the other methods do not exhibit a significant shift relative to SFT.

In Figure 4 (b), SFT+DPO and CRPO demonstrate comparable distributional shifts, which is the nature of DPO training that generally increases the reward of both chosen and rejected responses. Taken together, these findings indicate that the major reward difference achieved by CRPO, as observed in Figure 2, is primarily driven by elevating the rewards of desirable responses rather than solely relying on penalizing those of unfavored ones. This further substantiates our claim in Section 5.2 that the advantage of CRPO stems from promoting high-quality responses rather than solely from suppressing the probability of disfavored responses.

F Comparison with Another Ranked Optimization

To investigate whether the effectiveness of CRPO can be attributed simply to the use of ranked preference structure, we further compare CRPO with PRO (Song et al., 2024), which constructs preferences sequentially from a ranked set of responses:

$$\mathcal{L}_{\text{PRO}} = -\log \prod_{k=1}^{n-1} \frac{\exp\left(r_\pi(x, y_k)\right)}{\sum_{i=k}^{n} \exp\left(r_\pi(x, y_i)\right)}, \quad (8)$$

which is further combined with an NLL objective. Table 7 reports the AlpacaEval results across three target languages and two backbone models.

Despite also performing ranking-based preference optimization, PRO substantially underperforms CRPO. An inspection of generations from the PRO-trained models shows that they generally retain the ability to respond in the language of the input. This suggests that PRO can achieve the primary objective of *language adaptation*, but does not translate this adaptation into a comparable improvement in response quality. In contrast, CRPO improves both target-language and English AlpacaEval performance by a large margin.

Lang	Model	Method	Target		English	
			WR	LC	WR	LC
zh	Llama-3-8B	PRO	24.97	24.24	37.27	36.85
		CRPO	**62.36**	**62.45**	**57.20**	**56.73**
	Mistral-7B	PRO	13.42	13.75	36.15	33.67
		CRPO	**63.10**	**62.83**	**57.08**	**50.60**
id	Llama-3-8B	PRO	23.11	24.79	44.47	43.60
		CRPO	**68.40**	**67.97**	**64.35**	**64.24**
	Mistral-7B	PRO	13.29	12.42	29.81	30.52
		CRPO	**68.19**	**66.81**	**61.86**	**59.96**
sw	Llama-3-8B	PRO	11.43	11.78	37.64	36.17
		CRPO	**62.17**	**61.84**	**60.06**	**60.53**
	Mistral-7B	PRO	6.58	5.11	10.62	12.15
		CRPO	**51.98**	**50.76**	**57.76**	**56.21**

Table 7: AlpacaEval results of PRO and CRPO under the same training budget.

We attribute this difference to how the ranked preference structure is exploited during optimization. PRO primarily constructs pairwise comparisons between a higher-ranked response and lower-ranked alternatives. CRPO instead exploits the relative positions of responses throughout the ranked list and incorporates ranking-aware objectives to place greater emphasis on preference violations that affect the top of the ranking. Consequently, CRPO provides a more fine-grained learning signal for distinguishing responses with different quality levels rather than merely separating preferred from less-preferred responses.

This distinction is particularly important in our setting because language adaptation itself can already be encouraged by the NLL objective: the model can learn to generate in the target language without necessarily improving the quality of the generated content. The ranking-aware objective of CRPO complements this signal by explicitly preserving and promoting high-quality responses according to their positions in the ranked list. The considerable gap between PRO and CRPO therefore indicates that the gains of CRPO do not arise merely from introducing ranked preference optimization. Rather, *how* the ranked preference information is incorporated into the optimization objective is critical, further supporting the design of CRPO.