Cross-Lingual Optimization for Language Transfer in Large Language Models

Jungseob Lee*, **Seongtae Hong***, **Hyeonseok Moon**, **Heuiseok Lim**†
Korea University
{omanma1928, ghdchlwls123, glee889, limhseok}@korea.ac.kr

Abstract

Adapting large language models to other languages typically employs supervised fine-tuning (SFT) as a standard approach. However, it often suffers from an overemphasis on English performance, a phenomenon that is especially pronounced in data-constrained environments. To overcome these challenges, we propose **Cross-Lingual Optimization (CLO)** that efficiently transfers an English-centric LLM to a target language while preserving its English capabilities. CLO utilizes publicly available English SFT data and a translation model to enable cross-lingual transfer. We conduct experiments using five models on six languages, each possessing varying levels of resource. Our results show that CLO consistently outperforms SFT in both acquiring target language proficiency and maintaining English performance. Remarkably, in low-resource languages, CLO with only 3,200 samples surpasses SFT with 6,400 samples, demonstrating that CLO can achieve better performance with less data. Furthermore, we find that SFT is particularly sensitive to data quantity in medium and low-resource languages, whereas CLO remains robust. Our comprehensive analysis emphasizes the limitations of SFT and incorporates additional training strategies in CLO to enhance efficiency.

1 Introduction

While the rapid advancement of Large Language Models (LLMs) has led to significant innovations in natural language processing (Achiam et al., 2023; Waisberg et al., 2023), these models are primarily pre-trained on English data and consequently exhibit relatively lower performance for other languages (Touvron et al., 2023a; Dubey et al., 2024; Team et al., 2024). This discrepancy largely stems from the data distribution imbalance and scarcity of the pre-training and alignment data (Muraoka et al., 2023).

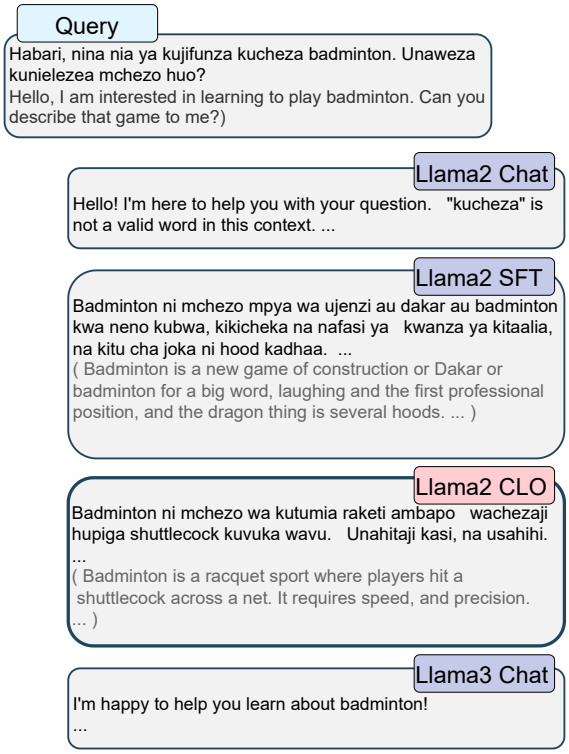

Figure 1: Example responses to a Swahili query generated by English-centric instruction models, the SFT model, and the proposed CLO model.

We identify the limited multilingual capabilities of English-centric language models in the following two cases. First, these models may fail to comprehend certain languages. Second, they might understand a language but still default to communicating in English (Marchisio et al., 2024). Figure 1 illustrates these scenarios. When posed with the Swahili question, the Llama2 Chat model (Touvron et al., 2023a) fails to comprehend Swahili adequately, while the Llama3 Chat model (Dubey et al., 2024) understands the query but is unable to generate responses in Swahili, opting for English instead. We notice that even the fine-tuned

*Equal contributor
†Corresponding author

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15100–15119
July 27 - August 1, 2025

model, despite being trained with 6,400 Swahili data, still struggles to produce appropriate outputs in Swahili (Zhao et al., 2024a; Chirkova and Nikoulina, 2024).

Based on these observations, we identified that standard fine-tuning (SFT) methods struggle to achieve direct alignment with English in data scarcity scenario. Then we hypothesize that simultaneously enhancing target language ability and aligning it with English will facilitate efficient transfer. In this context, we suggest a cross-lingual response prioritization method to strengthen the target language ability while aligning English and the target language. By promoting the preference for responding in the target language when provided with inputs in that same language, we argue that it is possible to utilize the model's embedded language knowledge while preserving its existing English capabilities.

To address this challenge, we propose the Cross-Lingual Optimization (CLO) strategy. CLO aims to effectively transfer an LLM to a target language using translated data. Specifically, it modifies the Direct Preference Optimization (DPO) (Rafailov et al., 2024) approach to increase the preference for responding in the same language as the input. Simultaneously, it reduces the preference for responding in different languages for a given input, thereby facilitating knowledge acquisition in the target language.

To validate the effectiveness of CLO, we experiment with a resources-limited environment. In this process, we used 6,400 publicly available English seed data points and an accessible translation model to target languages, along with five LLMs. Through comparative experiments with traditional transfer methods, we demonstrate superiority of the CLO method in terms of instruction-following ability and NLP benchmarks. Our results show that SFT exhibits better target language adaptation in high-resource settings, such as Chinese, while in low-resource languages like Swahili it tends to over-prioritize English. In contrast, our CLO approach demonstrated consistent performance improvements across all languages and models in both the target language and English. Furthermore, motivated by recent findings, we explored targeted fine-tuning strategies that reliably matched the performance of full model training. Accordingly, we adopt this training method as our primary strategy for facilitating effective language adaptation.

2 Related Works

Research on enhancing performance by transferring English-centric pre-trained language models to other languages is actively ongoing (Tran, 2020; Minixhofer et al., 2021; De Vries et al., 2021; Li et al., 2024; Chen and Lee, 2024). These studies primarily explore how effectively models trained in English can transfer with minimal data (Dobler and De Melo, 2023), as well as their ability to follow instructions in other languages (Zhao et al., 2024a).

Most language transfer studies either train English-centric LLMs with instruction tuning data in the target language (Lee et al., 2023; Shaham et al., 2024a) or perform instruction tuning after continual pre-training on a large corpus in the target language (Cui et al., 2023; Zhao et al., 2024a). However, many of these language transfer studies often require large-scale training datasets (Chen and Lee, 2024; Li et al., 2024) or involve complex model architecture analyses (Zhao et al., 2024b; Lee et al., 2024).

Inspired by Shaham et al. (2024b), which demonstrated high cross-lingual performance with a limited number of multilingual examples, we explore strategies for achieving effective transfer with small-sized datasets. In this paper, we propose an efficient language transfer method that addresses real-world scenarios, such as those in low-resource languages where instruction tuning data is unavailable, by leveraging publicly available translation models as a component and adapting an English-centric LLM to the target language using only English instruction tuning data.

3 Our Frameworks

Our cross-lingual transfer method assumes the availability of a base language model, a small amount of unidirectional English SFT data, and a translation model that supports the target language. The first key hypothesis is that given an input query in a non-English target language, suppressing English responses while strengthening responses in the target language enables the model to leverage its existing English knowledge to generate outputs in the target language. The second key hypothesis is that to transfer the ability to generate responses in English to the target language, it is sufficient to include a relatively small number of target language responses in a consistent response format.

Based on these hypotheses, we modify the Di-

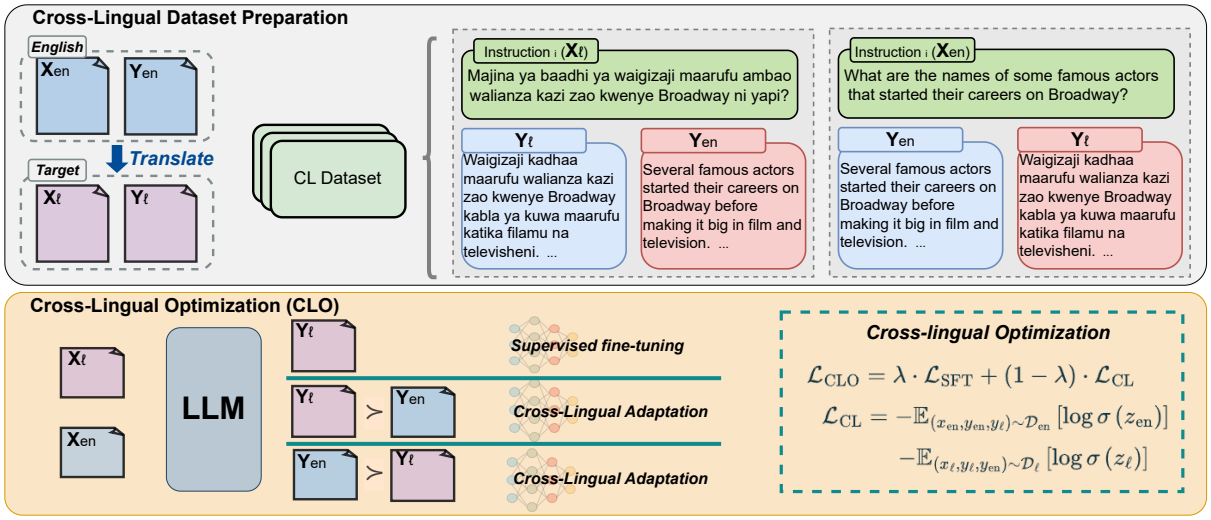

Figure 2: Overview of cross-lingual dataset preparation and optimization method. The process begins with translating English $(x_{\text{en}}, y_{\text{en}})$ pairs into a target language to create a cross-lingual dataset. This process results in the creation of (x_ℓ, y_ℓ) pairs in the target language. The optimization is performed using a combined loss $\mathcal{L}_{\text{CLO}}$.

rect Preference Optimization (DPO) loss (Rafailov et al., 2024) to suit our purpose. Since the base language model does not initially answer the queries, we combine the Negative Log-Likelihood (NLL) with the modified DPO loss to produce appropriate responses. Our process consists of two steps:

1. **Cross-Lingual Dataset Preparation**: Obtain translated data from the original English preference dataset.

2. **Cross-Lingual Optimization**: Train the attention layers using the translated data to generate responses in the query language.

Figure 2 illustrates our approach, and we detail the proposed CLO method below.

3.1 Brief Overview of Standard DPO

DPO employs an analytical mapping from a reward function to derive an optimal policy without the need for an explicit reward model. The standard DPO loss is defined as:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(\mathbf{x},\mathbf{y}_w,\mathbf{y}_l)\sim\mathcal{D}}\left[\log\sigma\left(z\right)\right] \quad (1)$$

where z is defined as:

$$z = \beta\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta\log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \quad (2)$$

where π_θ is the policy, π_{ref} the reference policy, σ represents the sigmoid function, and β regulates the KL constraint.

3.2 CLO Methodology

Cross-Lingual Dataset Preparation We generate target language prompts $\mathbf{x}_\ell$ and responses $\mathbf{y}_\ell$ by translating the English prompts $\mathbf{x}_{\text{en}}$ and responses $\mathbf{y}_{\text{en}}$ in the existing SFT dataset using a translation model.

For $\mathbf{x}_{\text{en}}$, we associate the *chosen response* with $\mathbf{y}_{\text{en}}$ and the *rejected response* with $\mathbf{y}_\ell$ to form the English training data. This mapping prevents the model from generating target language outputs when provided an English prompt, thereby preserving its English capability. Conversely, for $\mathbf{x}_\ell$, we map the *chosen response* to $\mathbf{y}_\ell$ and the *rejected response* to $\mathbf{y}_{\text{en}}$ to form the target language data. This encourages the model, when given a target language instruction, to generate responses in the target language by relying on its underlying English knowledge.

Cross-Lingual Optimization Our proposed CLO introduces a new loss function that differs from the standard DPO loss by utilizing cross-lingual data pairs within the *same* batch to explicitly teach the model the correspondence between input and output languages. Importantly, to mitigate inherent English bias, we only consider the NLL loss computed on the target language outputs. Further, Based on the findings of Zeping and Sophia (2024), which highlight the critical role of attention layers in language capabilities, we accordingly fine-tune only the attention layers. The

overall loss function of CLO is defined as:

$$\mathcal{L}_{\text{CLO}} = \lambda \cdot \mathcal{L}_{\text{SFT}} + (1 - \lambda) \cdot \mathcal{L}_{\text{CL}} \quad (3)$$

where $\mathcal{L}_{\text{SFT}}$ is the standard NLL loss over the target language data $\mathcal{D}_\ell$:

$$\mathcal{L}_{\text{SFT}} = \mathbb{E}_{(\mathbf{x}_\ell, \mathbf{y}_\ell) \sim \mathcal{D}_\ell} \left[-\log \pi_{\theta_{\text{att}}}(y_\ell \mid x_\ell) \right] \quad (4)$$

and $\mathcal{L}_{\text{CL}}$ is the cross-lingual loss that explicitly enforces language correspondence, defined as:

$$\begin{aligned} \mathcal{L}_{\text{CL}} = &- \mathbb{E}_{(\mathbf{x}_{\text{en}}, \mathbf{y}_{\text{en}}, \mathbf{y}_\ell) \sim \mathcal{D}_{\text{en}}} \left[\log \sigma \left(z_{\text{en}} \right) \right] \\ &- \mathbb{E}_{(\mathbf{x}_\ell, \mathbf{y}_\ell, \mathbf{y}_{\text{en}}) \sim \mathcal{D}_\ell} \left[\log \sigma \left(z_\ell \right) \right] \end{aligned} \quad (5)$$

where z_{en} and z_ℓ are defined respectively as:

$$z_{\text{en}} = \beta \left(\log \frac{\pi_{\theta_{\text{att}}}(y_{\text{en}} \mid x_{\text{en}})}{\pi_{\text{ref}}(y_{\text{en}} \mid x_{\text{en}})} - \log \frac{\pi_{\theta_{\text{att}}}(y_\ell \mid x_{\text{en}})}{\pi_{\text{ref}}(y_\ell \mid x_{\text{en}})} \right) \quad (6)$$

$$z_\ell = \beta \left(\log \frac{\pi_{\theta_{\text{att}}}(y_\ell \mid x_\ell)}{\pi_{\text{ref}}(y_\ell \mid x_\ell)} - \log \frac{\pi_{\theta_{\text{att}}}(y_{\text{en}} \mid x_\ell)}{\pi_{\text{ref}}(y_{\text{en}} \mid x_\ell)} \right) \quad (7)$$

By incorporating both preferred and rejected responses in different languages within the same batch, the model is guided to generate outputs in the appropriate language based on the input, effectively transferring its English knowledge to the target language while maintaining proficiency in English. The differences between the standard DPO loss and the Cross-lingual loss are detailed in Appendix A.

4 Experimental Setup

4.1 Training Setup

We primarily compare our CLO method with the standard SFT baseline. In some experiments, we also include results for models trained with SFT followed by DPO, following the experimental protocol described by Hong et al. (2024). Specifically, *SFT+DPO* refers to a two-stage training procedure in which SFT is first performed using the preferred responses for each data point, and then DPO is applied to the same dataset. This approach evaluates the effect of simple data augmentation using cross-lingual datasets. The experimental results indicate that using cross-lingual datasets without the CLO objective function does not achieve strong language transfer performance.

We use 6,400 data samples from OpenAssistant (Köpf et al., 2024), selected in order of highest ranking[1], as our English single-turn dataset. Each example is a selected (instruction, output) pair from the first turn.

To validate the effectiveness of CLO, we conduct experiments on six languages with varying resource availability, using five available pretrained LLMs: Llama-2-7B, Llama-2-13B (Touvron et al., 2023b), Llama-3-8B (Dubey et al., 2024), Mistral-7B-v0.1 (Jiang et al., 2023), and Qwen-2.5-3B (Yang et al., 2024). Specifically, we choose two high-resource languages (Chinese, German), two medium-resource languages (Korean, Indonesian), and two low-resource languages (Swahili, Yoruba) based on the amount of pretraining corpus data in the Llama-2 (Touvron et al., 2023b). For each of the specified target languages, we create cross-lingual data by translating all the extracted English (instruction, output) pairs $\{(\mathbf{x}_{\text{en}}, \mathbf{y}_{\text{en}})\}$ into $\{(\mathbf{x}_{\text{en}}, \mathbf{y}_{\text{en}}, \mathbf{x}_\ell, \mathbf{y}_\ell)\}$.

We adopt the M2M100 1.2B translation model (Fan et al., 2021) to generate a total of 12,800 samples across the languages (6,400 in English and 6,400 in the target language). We trained all baselines and models using identical hyperparameters and conducted evaluations under the same settings. The details of the hyperparameters used during training and generation are provided in Appendix B.

4.2 Evaluation Benchamrk Setup

AlpacaEval To investigate the instruction-following ability of the models, we use AlpacaEval (Li et al., 2023), translating it into each respective language, and measure the win rate against the baseline SFT model. We include in the evaluation whether the model responded appropriately in the given language, in addition to the original prompts from AlpacaEval, and use the GPT-4o[2] for comparisons. The evaluation prompts and detailed information are provided in Appendix C.

MRC Benchmark To assess the machine reading comprehension (MRC) capabilities of the models, we employ the BELEBELE dataset (Bandarkar et al., 2023). In this benchmark, each instance consists of a question and a passage accompanied by four answer choices. Each instance features human-curated questions and answers that are designed to be challenging by including plausible distractors. The model's accuracy is determined by computing the log likelihood of each answer op-

[1] The dataset provides a `rank` column, which indicates the relative quality or preference order of assistant responses to a given prompt, as judged by human annotators.

[2] https://openai.com/index/hello-gpt-4o/, *gpt-4o-2024-08-06*

Model	Eval Lang	High-Resource						Medium-Resource						Low-Resource					
		Chinese			German			Korean			Indonesian			Swahili			Yoruba		
		SFT+DPO	CLO	Δ	SFT+DPO	CLO	Δ	SFT+DPO	CLO	Δ	SFT+DPO	CLO	Δ	SFT+DPO	CLO	Δ	SFT+DPO	CLO	Δ
Llama-3-8B	Target	59.1±1.70	70.4±1.62	**+11.3**	51.7±1.76	54.6±1.76	**+2.9**	62.5±1.70	77.8±1.47	**+15.3**	54.9±1.75	56.4±1.75	**+1.5**	65.4±1.67	83.0±1.61	**+17.6**	52.4±1.76	64.0±1.70	**+11.6**
	English	52.0±1.76	64.0±1.73	**+12.0**	50.1±1.76	55.5±1.76	**+5.4**	52.1±1.76	64.4±1.69	**+12.3**	51.6±1.76	57.7±1.75	**+6.1**	50.3±1.76	61.4±1.73	**+11.1**	51.2±1.76	58.2±1.75	**+7.0**
Llama-2-7B	Target	59.1±1.73	61.1±1.72	**+2.0**	50.3±1.76	59.5±1.73	**+9.2**	52.5±1.76	53.8±1.76	**+1.3**	50.8±1.76	55.8±1.75	**+5.0**	64.1±1.64	65.0±1.72	**+0.9**	43.5±1.74	67.1±1.66	**+23.6**
	English	51.1±1.76	55.7±1.76	**+4.6**	48.4±1.76	61.2±1.72	**+12.8**	50.6±1.76	51.0±1.76	**+0.4**	49.3±1.76	62.4±1.71	**+13.1**	51.5±1.76	55.5±1.75	**+4.0**	50.3±1.76	61.2±1.72	**+10.9**
Llama-2-13B	Target	59.3±1.74	65.2±1.71	**+5.9**	50.5±1.76	53.7±1.76	**+3.2**	51.6±1.75	53.9±1.75	**+2.3**	52.4±1.76	61.7±1.71	**+9.3**	53.9±1.56	70.9±1.60	**+17.0**	43.5±1.75	67.3±1.65	**+23.8**
	English	50.6±1.76	50.4±1.76	**-0.2**	53.0±1.76	58.5±1.74	**+5.5**	51.4±1.76	52.3±1.76	**+0.9**	52.9±1.76	59.8±1.73	**+6.9**	50.8±1.76	55.0±1.75	**+4.2**	51.1±1.76	54.2±1.76	**+3.1**
Mistral-7B-v0.1	Target	56.8±1.75	57.4±1.74	**+0.6**	49.9±1.76	50.8±1.76	**+0.9**	48.5±1.76	50.5±1.76	**+2.0**	50.0±1.76	51.1±1.76	**+1.1**	35.4±1.71	51.3±1.76	**+15.9**	50.9±1.76	51.1±1.52	**+0.2**
	English	51.4±1.76	57.1±1.75	**+5.7**	48.6±1.76	50.6±1.76	**+2.0**	52.9±1.76	65.2±1.71	**+12.3**	52.3±1.76	54.8±1.76	**+2.5**	51.2±1.76	64.2±1.69	**+13.0**	49.4±1.76	52.8±1.76	**+3.4**
Qwen-2.5-3B	Target	54.2±1.75	59.9±1.73	**+5.7**	53.0±1.76	54.0±1.76	**+1.0**	52.3±1.76	62.4±1.71	**+10.1**	53.8±1.76	54.9±1.75	**+1.1**	51.7±1.76	74.7±1.53	**+23.0**	45.9±1.76	68.9±1.63	**+23.0**
	English	50.7±1.76	56.7±1.75	**+6.0**	51.9±1.76	54.8±1.76	**+2.9**	50.8±1.76	61.0±1.72	**+10.2**	50.6±1.76	58.0±1.75	**+7.4**	51.1±1.76	58.1±1.74	**+7.0**	54.4±1.76	57.0±1.75	**+2.6**

Table 1: Win-rate (%) results on AlpacaEval for models fine-tuned with SFT+DPO and CLO, evaluated against their SFT baselines. Each cell reports the win rate and its standard deviation. The Δ denotes the absolute improvement of CLO over SFT+DPO.

tion and selecting the one with the highest overall likelihood (Gao et al., 2024).

To support our multilingual experiments, we use language-specific prompt templates that maintain a consistent structure across languages while using native terms for key input fields. In these templates, the input fields are mapped to their corresponding local labels. Table 7 describes these mappings for each target language, including the appropriate answer indicator. In our experiments, each formatted prompt is passed as a single text string into the model to evaluate its zero-shot performance across six languages.

MMMLU To measure the models' reasoning performance, we utilize OpenAI's Multilingual Massive Multi-task Language Understanding[3] (MMMLU) (Hendrycks et al., 2021). This test set, developed with the input of professional human translators, extracts answers not based on the probabilities of the correct tokens but from the model's generated outputs. Since the model infers answers based on the step-by-step reasoning path, it allows for more accurate measurements than traditional token probabilities. Additionally, because the test includes instructions that require the model to respond in specific answer formats, the model's instruction-following ability is essential. For all MMMLU experiments, we simply perform a single run with each model to compute the score, without aggregating results from multiple runs.

Furthermore, we evaluate MMMLU using language-specific prompts[4] and, to assess the impact of interference when integrating target language data, we include two SFT variants: one trained exclusively on English data (*SFT-eng*) and another trained only on target language data (*SFT-tgt*). We conduct these experiments on five models across three languages.

5 Experimental Results

We present the experimental results on three key benchmarks: AlpacaEval, which measures instruction following ability, BELEBLE, which evaluate the models' MRC abilities, and MMMLU, which assesses the reasoning performance.

Instruction Following Ability We evaluate the instruction following ability of the models using AlpacaEval, and the results are presented in Table 1.[5] The SFT+DPO enhances the performance of the SFT baseline in target languages, maintaining comparable performance in English for high-resource languages. However, this configuration yields only modest improvements in medium-resource languages and demonstrates a tendency to prioritize English output in low-resource languages. Conversely, the CLO consistently surpasses the SFT across all base models and languages, achieving a win rate exceeding 50% in all cases. CLO demonstrates substantial advancements in medium-resource languages, indicative of its consistent alignment with the target language. Even though CLO outperforms SFT in English, SFT+DPO only partially enhances SFT and fails to close the performance gap with CLO. This discrepancy is particularly pronounced in low-resource languages, where CLO significantly outperforms SFT+DPO, a result possibly attributable to the detrimental effect of excessive parameter fluctuations during fine-tuning due to insufficient embedded language knowledge.

[3] https://github.com/openai/simple-evals

[4] In the original instructions are provided in English. To ensure accurate linguistic evaluation, we use the language-specific prompts. Detailed MMMLU test settings and the modified instructions are provided in Appendix D.

[5] Following Marchisio et al. (2024), we report additional results for the target language evaluation in Table 8, wherein the instruction 'Please answer in the same language as the input' is translated into the target language and appended to the original prompt.

Model	Method	High-Resource				Medium-Resource				Low-Resource			
		Chinese		German		Korean		Indonesian		Swahili		Yoruba	
		Target	English	Target	English	Target	English	Target	English	Target	English	Target	English
Llama-2-7B	SFT	36.0±1.60	35.6±1.61	31.7±1.55	36.9±1.61	27.4±1.49	37.1±1.61	30.0±1.53	35.8±1.60	23.9±1.42	36.2±1.60	26.0±1.46	36.2±1.60
	CLO	**36.7**±1.61	**37.1**±1.60	**32.7**±1.56	**38.1**±1.62	**30.1**±1.53	**37.9**±1.62	**31.7**±1.55	**36.6**±1.61	**28.6**±1.51	**37.3**±1.61	**29.0**±1.51	**37.6**±1.62
Llama-2-13B	SFT	48.2±1.67	58.6±1.64	51.1±1.67	59.6±1.64	37.9±1.62	53.6±1.66	46.7±1.66	60.9±1.63	26.6±1.47	57.4±1.65	25.1±1.45	60.3±1.63
	CLO	**51.3**±1.67	**58.8**±1.64	**52.3**±1.67	**59.8**±1.64	**38.7**±1.62	**57.6**±1.65	**47.8**±1.67	**61.3**±1.62	**32.3**±1.56	**58.7**±1.64	**27.2**±1.48	**60.7**±1.63
Llama-3-8B	SFT	69.9±1.53	76.3±1.42	62.9±1.61	73.1±1.45	46.7±1.66	64.9±1.59	57.0±1.65	**75.2**±1.44	42.0±1.64	75.0±1.44	29.6±1.52	76.0±1.42
	CLO	**70.6**±1.52	**77.1**±1.40	**64.1**±1.60	**74.7**±1.48	**57.7**±1.65	**73.4**±1.47	**58.7**±1.64	75.0±1.44	**42.6**±1.65	**75.9**±1.43	**29.8**±1.53	**76.3**±1.42
Mistral-7B-v0.1	SFT	56.3±1.63	70.6±1.52	47.6±1.67	67.9±1.56	29.2±1.52	**70.6**±1.52	51.1±1.67	72.3±1.49	34.3±1.58	**72.3**±1.49	31.7±1.55	68.2±1.55
	CLO	**61.1**±1.65	**70.7**±1.52	**53.0**±1.66	**72.3**±1.49	**48.0**±1.67	58.8±1.64	**53.4**±1.66	**75.7**±1.43	**38.8**±1.63	69.8±1.53	**31.9**±1.55	**74.7**±1.45
Qwen2.5-3B	SFT	83.1±1.25	81.6±1.29	70.3±1.52	81.3±1.30	67.7±1.56	80.9±1.31	65.3±1.59	81.8±1.29	36.6±1.61	82.9±1.25	27.0±1.48	80.9±1.29
	CLO	**83.4**±1.24	**82.3**±1.27	**73.1**±1.48	**82.0**±1.28	**68.8**±1.55	**82.0**±1.28	**67.1**±1.57	**83.1**±1.25	**39.8**±1.63	**83.1**±1.26	**29.6**±1.52	**81.7**±1.31

Table 2: Zero-shot evaluation results for BELEBELE. In each cross-lingual setting, "Target" refers to accuracy for the target language, while "English" indicates accuracy for English. The standard deviations (±) are also reported.

Remarkably, the Llama-3, with its extensive internal knowledge, and Qwen-2.5, known for its outstanding multilingual capabilities, exhibit the greatest performance gains when using CLO, highlighting CLO's efficacy in leveraging internal English knowledge for enhanced cross-lingual alignment with the target language. These findings suggest that SFT training may overly prioritize English, with target language data inducing disruptive parameter fluctuations that further undermine the model's multilingual capabilities. In contrast, CLO offers a more equitable adaptation approach, sustaining robust English performance while adeptly accommodating the target language. Moreover, the results evaluated using the original AlpacaEval prompt, without our language-specific prompts, are presented in Appendix G, with outcomes generally showing similar performance in the target language. Upon manual review of the differently evaluated results, we observed discrepancies only when responses were provided in another language at the word level.[6] These observations suggest that our language-specific prompt is adequately assessing the model's capability in handling the target language.

MRC Performances The results are reported in Table 2. Overall, CLO exhibits superior performance compared to SFT by effectively adapting target language. By contrast, SFT either struggles to adapt fully to the target languages or remains overly reliant on English, resulting in decreased accuracy.

Notably, Qwen2.5, despite having the smallest parameter scale among the compared models, demonstrates particularly strong performance in Chinese. We believe that this is because Qwen2.5 was pre-trained on a substantial amount of Chinese data, facilitating its adaptation to Chinese-specific tasks. Furthermore, Qwen2.5 also achieves high accuracy in the other target languages, suggesting that its multilingual pre-training facilitates robust cross-lingual adaptation (Yang et al., 2024).

Reasoning Performance We evaluate the models on MMMLU, with the results summarized in Table 3.[7] The proposed CLO generally demonstrates higher performance across most of the languages and models tested. The CLO Llama-3-8B models exhibits outstanding performance, achieving higher scores compared to the SFT. This underscores the significant enhancements gained through leveraging the model's extensive internal knowledge via the CLO method. In contrast, the SFT tends to show relatively high performance in English but lower performance in target languages, especially in low-resource settings. This suggests that traditional SFT methods may not adapt well when limited data is available and tend to learn in an English-centric manner. While applying DPO to SFT models using cross-lingual datasets can generally improve performance, we observe that this approach is cost-inefficient and does not consistently guarantee enhanced performance in target languages. In some cases, SFT+DPO still exhibit English-centric learning patterns.

We observed that, except for the Mistral, the performances of the *SFT-eng* model are similar to that of the SFT models trained on both languages. These results suggest that in high-resource languages like Chinese, there is no loss in English performance when additional language data is in-

[6]Excluding the Mistral, no word-level confusion is observed; hence, we do not provide additional Word-level Confusion analysis (Marchisio et al., 2024).

[7]Most models experience extraction failures in less than 0.8% of cases, indicating that such failures have minimal impact on performance evaluation.

Model	Method	Chinese		Korean		Swahili	
		Target	English	Target	English	Target	English
Llama-2-7B	SFT-eng	–	29.40	–	29.40	–	29.40
	SFT-tgt	27.19	–	25.31	–	22.46	–
	SFT	27.12	29.78	23.47	26.01	17.21	28.99
	SFT+DPO	26.59	31.00	28.00	29.88	19.96	28.10
	CLO	**28.11**	**31.24**	**29.09**	**31.54**	**24.09**	**30.14**
Llama-2-13B	SFT-eng	–	41.81	–	41.81	–	41.81
	SFT-tgt	31.51	–	36.80	–	22.56	–
	SFT	31.17	42.47	34.39	37.69	21.38	36.77
	SFT+DPO	33.26	43.75	26.79	40.22	19.83	**46.32**
	CLO	**34.17**	**47.20**	**39.70**	**44.14**	**26.78**	41.68
Llama-3-8B	SFT-eng	–	49.13	–	49.13	–	49.13
	SFT-tgt	38.55	–	29.61	–	29.05	–
	SFT	39.36	53.00	25.31	50.61	27.59	44.48
	SFT+DPO	40.91	56.36	27.48	50.71	28.86	**55.97**
	CLO	**41.99**	**57.55**	**32.73**	**55.57**	**33.38**	55.82
Mistral-7B	SFT-eng	–	50.67	–	50.67	–	50.67
	SFT-tgt	33.16	–	27.65	–	28.61	–
	SFT	33.74	52.11	25.94	34.11	26.68	49.95
	SFT+DPO	**37.02**	52.03	26.77	36.73	**28.42**	**53.39**
	CLO	34.05	**58.74**	**28.31**	**50.95**	28.07	50.73
Qwen2.5-3B	SFT	46.34	55.70	35.90	56.01	26.22	57.19
	CLO	**52.10**	**60.52**	**41.94**	**61.49**	**29.80**	**60.92**

Table 3: MMMLU evaluation results. The performance of *CLO*, *SFT*, *SFT+DPO*, *SFT-eng*, and *SFT-tgt* models trained on CLO datasets in Chinese, Korean, and Swahili across four base models. *SFT-eng* indicates the performance of models where SFT is performed using only English data, while *SFT-tgt* denotes the performance of models where SFT is performed using only target language data.

cluded. The inclusion of the Chinese dataset appears to act as data augmentation, enhancing the model's robustness and resulting in better performance compared to training on English data alone. In contrast, for medium-resource languages like Korean (excluding the Llama-3), *SFT-eng* outperforms the SFT model trained on both English and Korean. This discrepancy becomes even more pronounced in the low-resource setting of Swahili. These findings indicate that incorporating medium-resource Korean and low-resource Swahili data can potentially degrade the model's English capabilities. We speculate that in the case of the Llama-3, Korean is also learned with relatively high-resource data, which might explain why training on both languages yields better performance compared to the traditional SFT approach. This hypothesis is further supported by the performance of the *SFT-tgt*. In the high-resource Chinese setting, the *SFT-tgt* model's performance is comparable to that of the SFT model trained on both datasets. However, in medium-resource Korean and low-resource Swahili, the *SFT-tgt* model records higher performance than the SFT, providing clearer evidence to support our speculation.

Additionally, since Llama-3-8b exhibited the most pronounced changes in previous experiments, we conducted a category-wise analysis of its MMMLU performance in the target languages, as presented in Appendix H. In conclusion, the CLO Llama-3 demonstrated strong performance across a wide range of categories, whereas the SFT model showed decreased performance in specialized domains, highlighting the limitations of the SFT method in low-data scenarios.

6 Effect of Training Data Size

To evaluate the impact of training data size on the performance of CLO and SFT, we conduct experiments by varying the amount of training data. Specifically, we adjust the number of training examples per language to 200, 400, 800, 1600, 3200, and 6400 based on single-turn English data. For comparison, we utilize the AlpacaEval test setup employed, and each model trained with a different data size is compared against the SFT model trained on 6,400 pair examples. The experimental results of Llama-2-7b are depicted in Figure 3 and the results for Llama-3-8b are presented in Appendix I.

The results of Llama-2 reveal that both CLO and SFT achieve relatively rapid performance improvements with smaller amounts of data in Chinese, a high-resource language. However, for SFT, performance improvements in Korean, a medium-resource language, are observable only when the

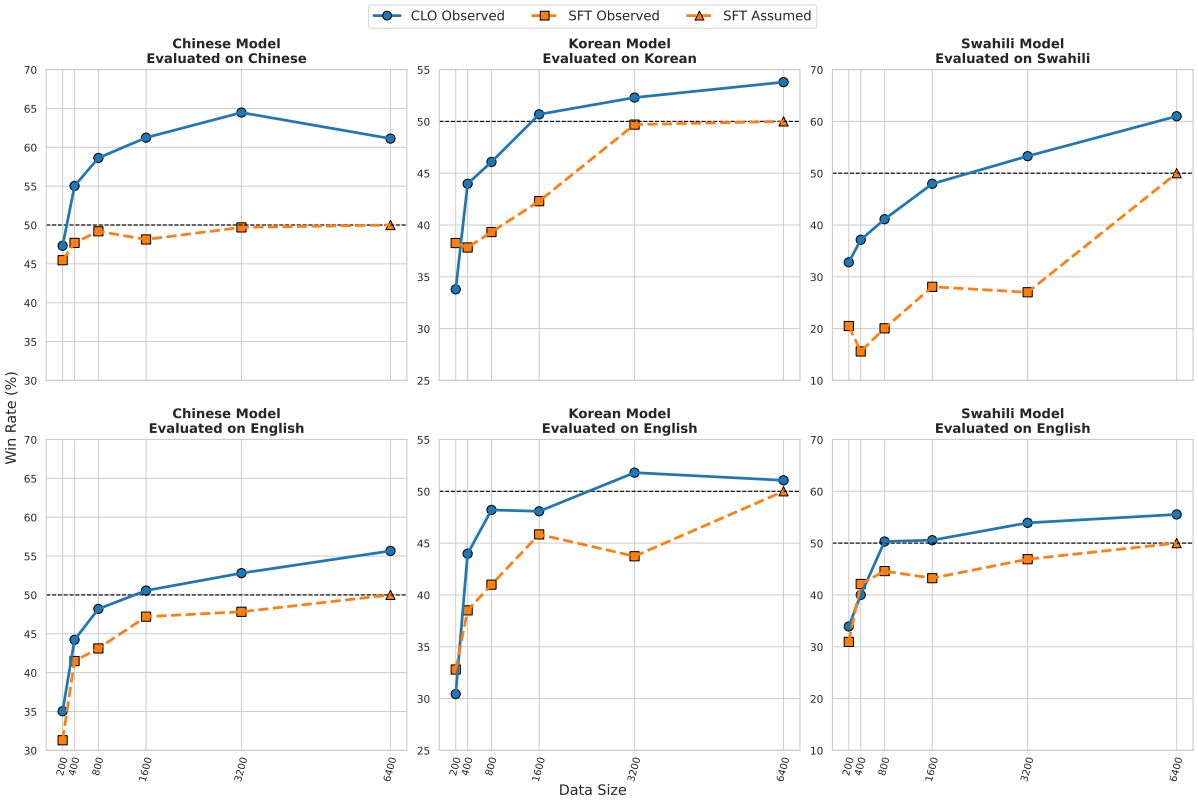

Figure 3: Comparison of win rates between CLO and SFT on Llama-2-7B models trained with varying amounts of data, evaluated against a SFT with 6,400 pair examples on the AlpacaEval. The 'SFT Assumed' baseline is assigned a win rate of 50%, as it compares identical models.

training data exceeds 3,200 pairs. Moreover, in Swahili, a low-resource language, there is a significant performance difference between models trained with 3200 and 6,400 pairs. In contrast, CLO demonstrates efficient performance enhancement across all languages, even in low-data environments. Notably, in Swahili, CLO exhibits exceptional performance by achieving results comparable to the SFT model trained on 6,400 pairs using only 1600 pairs.

We find that in Llama-3, the SFT method achieves similar performance with only 1,600 pairs as it does with 6,400 pairs for high-resource language Chinese and medium-resource language Korean. However, for the low-resource language Swahili, the SFT method requires 6,400 pairs to achieve comparable performance. On the other hand, the CLO method attains performance comparable to the SFT model trained on 6,400 pairs using only 400 pairs in the target language. Both the results of Llama-2 and Llama-3 indicate that transfer to low-resource languages is more challenging for the SFT than for the CLO.

Overall, our findings confirm that CLO attains faster performance improvements compared to SFT in Chinese, Korean, and Swahili. This demonstrates that CLO can facilitate rapid language transfer even with relatively small amounts of data. Particularly in Swahili, CLO achieves performance similar to the SFT model trained on 6,400 pairs using merely 1,600 pairs. In summary, CLO enhances performance more efficiently than SFT across diverse language environments. Especially in low-resource languages, CLO outperforms the existing SFT models with substantially less training data. Conversely, SFT exhibits a strong dependence on the quantity of training data in medium and low-resource languages, indicating potential limitations in performance improvement when sample data collection is infeasible.

7 Ablation Studies

7.1 Comparison with Full Tuning

According to Zeping and Sophia (2024), certain types of knowledge related to language can be mostly stored in the attention layers of language models, and important neurons are concentrated in deeper layers. Based on these results, we applied a

Models	Target Language		English Evaluation	
	Win (%)	*Lose (%)*	*Win (%)*	*Lose (%)*
Chinese	**50.68**	49.07	**54.41**	45.59
Korean	**52.73**	46.40	**52.67**	47.33
Swahili	29.69	**69.57**	**50.06**	49.94

Table 4: Comparison of AlpacaEval generation performance between the only trained attention Llama-2 CLO ($\pi_{\theta_{\text{attn}}}$) and the all parameter trained Llama-2 CLO (π_θ) models trained on Chinese, Korean, and Swahili.

NLL Loss	Swahili	English
Target-Only (ours)	83.0	65.4
Target & English	74.5	67.8

Table 5: Comparison of win rates on the AlpacaEval dataset for SFT and CLO variants in Swahili, both trained with 6,400 pair examples on the Llama-3-8B model. Results are presented using only target language NLL loss and both target and English NLL losses.

method in CLO where only the attention layers are trained during parameter updates, and compared the performance with updating all parameters, as shown in Table 4.

Experimental results confirmed that updating only the attention layers maintains or even improves the model's performance. This indicates that in our method, language alignment for knowledge representation and utilization can be achieved by training only the attention layers, suggesting the potential of an efficient model update strategy. However, our results found that in Swahili, the attention-only training CLO method showed a significant performance drop compared to full training. This result suggests that aligning the language by training only the attention layers in low-resource languages like Swahili is challenging.

7.2 NLL Loss Analysis

Furthermore, we conducted ablation experiments to assess the effect of using a combined target and English NLL loss versus employing only the target language NLL loss, with the results presented in Table 5. The findings indicate that incorporating both losses inadvertently biases the model toward English responses, as evidenced by the higher performance on English at the expense of the target language. Consequently, we adopt the approach that considers only the target language NLL loss, which better maintains a balanced performance between the target and English languages.

8 Conclusion

In this paper, we introduced **Cross-Lingual Optimization (CLO)**, an effective strategy for transferring English-centric LLMs to target languages while preserving their English capabilities. Leveraging publicly available English SFT data and translation models, CLO facilitates cross-lingual transfer without the need for extensive target language data.

We conduct experiments using five LLMs across six languages with varying resource levels: two high-resource (Chinese, German), two medium-resource (Korean, Indonesian), and two low-resource (Swahili, Yoruba). Our results show that CLO outperforms SFT in both acquiring target language proficiency and maintaining English performance. Notably, in the low-resource language, CLO achieved superior results with only 3,200 pairs, surpassing SFT models trained on twice the amount of data. We found that traditional SFT is particularly sensitive to data quantity in medium and low-resource languages, often leading the model to either overly rely on its English knowledge or diminish it when data is scarce, resulting in insufficient adaptation to the target language. In contrast, CLO's approach for responding in the target language enables it to utilize embedded language knowledge more effectively, leading to better performance even with less data.

Limitations

Multilinguality This study focuses on the cross-lingual transfer of an English-centric large language model to a specific target language rather than expanding to multiple languages simultaneously. Consequently, our research does not address the enhancement of multilingual performance across several languages at once. This limitation suggests the need for future work to explore methods for transferring to multiple languages concurrently to improve overall multilingual capabilities.

Training Data Our approach relies on translated data to address the scarcity of human-produced target-language instruction data, particularly for low-resource languages. Translation models can introduce semantic distortions or uncertainties, and our current work does not explicitly quantify how these factors might affect the model's performance. However, CLO demands the use of datasets that are aligned between English and the target language,

which is a resource that is extremely costly and often impractical to curate manually at scale. Given these constraints, employing a translation model becomes a necessity, especially for low-resource languages.

Since both our baseline SFT and the proposed CLO method operate on the same translated training set, any translation artifacts or alignment issues are likely to impact both approaches uniformly, thus ensuring a fair comparison of their relative performance. Moreover, manual inspection suggests that major translation errors were minimal, and our empirical results indicate that CLO consistently outperforms SFT across language settings, suggesting that minor translation imperfections do not undermine the advantages of leveraging embedded English knowledge to enhance target-language capability.

Limited Scope of Languages Our study was limited to experiments involving English and six target languages (Chinese, German, Korean, Indonesian, Swahili, and Yoruba). We selected these six because extending the analysis to additional languages would require training and evaluating three separate models (SFT, SFT+DPO, and CLO), which would introduce significant additional costs and time constraints. Consequently, there are limitations in generalizing the results to all languages, and it is necessary to examine the potential for extending our methods to a wider variety of languages in future work.

Evaluation on Language-Specific Data Our experiments evaluated language models using AlpacaEval datasets translated by GPT-4o and the MMMLU dataset, which was translated into respective languages by professional human translators, correcting the initial misstatement. Additionally, we assessed Machine Reading Comprehension abilities using the BELEBELE (Bandarkar et al., 2023) dataset, constructed and reviewed by human translators. While these datasets are suitable for measuring general model performance, they do not fully capture the model's ability to respond appropriately to data involving linguistic characteristics or cultural contexts specific to each language. Consequently, we were unable to thoroughly evaluate the model's handling of language-specific nuances or culturally relevant content.

Applicability to Other Methods The proposed CLO method was experimented with only by applying it to the DPO Direct Preference Optimization methodology, although it can potentially be applied to various preference optimization algorithms (Ethayarajh et al., 2024; Hong et al., 2024; Xie et al., 2024). However, since our experiments did not study the generality of CLO across these different preference optimization methods, further research is needed to verify whether CLO can guarantee performance improvements in other methodologies.

Computational Efficiency The CLO method performs parameter updates based on a reference model and trains only a subset of the total parameters, resulting in a lower computational cost than DPO while incurring a slightly additional cost relative to SFT training. In our environment, the GPU memory allocation required for CLO is up to 55% higher than that of conventional SFT training, whereas it is approximately 30% percent lower than that required by DPO. Moreover, our method does not allow for an exact computation of FLOPs, making precise inference of the training cost difficult. In our experiments, the training time for CLO is nearly identical to or slightly higher than that of SFT.

Acknowledgements

This work was partly supported by ICT Creative Consilience Program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (IITP-2025-RS-2020-II201819, 25%), Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2021R1A6A1A03045425, 25%), 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, 25%), and Institute of Information & communications Technology Planning & Evaluation (IITP) under the Leading Generative AI Human Resources Development (IITP-2025-R2408111, 25%) grant funded by the Korea government (MSIT).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman,

Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*.

Kuang-Ming Chen and Hung-yi Lee. 2024. Instructioncp: A fast approach to transfer large language models into target language. *arXiv preprint arXiv:2405.20175*.

Nadezhda Chirkova and Vassilina Nikoulina. 2024. Zero-shot cross-lingual transfer in instruction tuning of large language model. *arXiv preprint arXiv:2402.14778*.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Wietse De Vries, Martijn Bartelds, Malvina Nissim, and Martijn Wieling. 2021. Adapting monolingual models: Data can be scarce when language similarity is high. *arXiv preprint arXiv:2105.02855*.

Konstantin Dobler and Gerard De Melo. 2023. Focus: Effective embedding initialization for monolingual specialization of multilingual models. *arXiv preprint arXiv:2305.14481*.

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*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

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

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

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*, 2(4):5.

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*.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. 2024. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36.

Jungseob Lee, Hyeonseok Moon, Seungjun Lee, Chanjun Park, Sugyeong Eo, Hyunwoong Ko, Jaehyung Seo, Seungyoon Lee, and Heui-Seok Lim. 2024. Length-aware byte pair encoding for mitigating oversegmentation in korean machine translation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2287–2303.

SeungJun Lee, Taemin Lee, Jeongwoo Lee, Yoonna Jang, and Heuiseok Lim. 2023. Kullm: Learning to construct korean instruction-following large language models. In *Annual Conference on Human and Language Technology*, pages 196–202. Human and Language Technology.

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`.

Yudong Li, Yuhao Feng, Wen Zhou, Zhe Zhao, Linlin Shen, Cheng Hou, and Xianxu Hou. 2024. Dynamic data sampler for cross-language transfer learning in large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11291–11295. IEEE.

Kelly Marchisio, Wei-Yin Ko, Alexandre Bérard, 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.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2021. Wechsel: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. *arXiv preprint arXiv:2112.06598*.

Masayasu Muraoka, Bishwaranjan Bhattacharjee, Michele Merler, Graeme Blackwood, Yulong Li, and Yang Zhao. 2023. Cross-lingual transfer of large language model by visually-derived supervision toward low-resource languages. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3637–3646.

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

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024a. Multilingual instruction tuning with just a pinch of multilinguality. *arXiv preprint arXiv:2401.01854*.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024b. Multilingual instruction tuning with just a pinch of multilinguality. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2304–2317, Bangkok, Thailand. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

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

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*.

Ke Tran. 2020. From english to foreign languages: Transferring pre-trained language models. *arXiv preprint arXiv:2002.07306*.

Ethan Waisberg, Joshua Ong, Mouayad Masalkhi, Sharif Amit Kamran, Nasif Zaman, Prithul Sarker, Andrew G Lee, and Alireza Tavakkoli. 2023. Gpt-4: a new era of artificial intelligence in medicine. *Irish Journal of Medical Science (1971-)*, 192(6):3197–3200.

Tengyang Xie, Dylan J Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed Awadallah, and Alexander Rakhlin. 2024. Exploratory preference optimization: Harnessing implicit q*-approximation for sample-efficient rlhf. *arXiv preprint arXiv:2405.21046*.

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

Yu Zeping and Ananiadou Sophia. 2024. Neuron-level knowledge attribution in large language models. In *Submitted to ACL Rolling Review - June 2024*. Under review.

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

Yiran Zhao, Wenxuan Zhang, Huiming Wang, Kenji Kawaguchi, and Lidong Bing. 2024b. Adamergex: Cross-lingual transfer with large language models via adaptive adapter merging. *arXiv preprint arXiv:2402.18913*.

A Details of Cross-Lingual Loss

Cross-Lingual Dataset Preparation We generate target language prompts $\mathbf{x}_\ell$ and target language responses $\mathbf{y}_\ell$ by translating the English prompts $\mathbf{x}_{\text{en}}$ and English responses $\mathbf{y}_{\text{en}}$ in the existing English preference dataset using a translation model.

For $\mathbf{x}_{\text{en}}$, we map the *chosen response* to $\mathbf{y}_{\text{en}}$ and the *rejected response* to $\mathbf{y}_\ell$, constructing the English training data. This is to prevent the model from responding in the target language when given an English prompt, thereby preserving the existing English knowledge within the model.

Conversely, for $\mathbf{x}_\ell$, we map the *chosen response* to $\mathbf{y}_\ell$ and the *rejected response* to $\mathbf{y}_{\text{en}}$, constructing the target language data. This is to suppress the model's tendency to respond in English when given a target language instruction, encouraging it instead to utilize its English knowledge to generate outputs in the target language.

Cross-Lingual Loss Function Our proposed CLO introduces a new loss function that differs from the standard DPO loss by utilizing cross-lingual data pairs within the *same batch* to explicitly teach the model the correspondence between input and output languages.

The overall objective function of CLO is defined as:

$$\mathcal{L}_{\text{CLO}} = \lambda \cdot \mathcal{L}_{\text{SFT}} + (1 - \lambda) \cdot \mathcal{L}_{\text{CL}} \tag{8}$$

Here, $\mathcal{L}_{\text{SFT}}$ is the supervised fine-tuning loss that promotes the language model to generate correct outputs for the target language only. It is calculated using the Negative Log-Likelihood (NLL) over the target-language data in the batch:

$$\mathcal{L}_{\text{SFT}} = \frac{1}{N} \sum_{i=1}^{N} \left[-\log \pi_{\theta_{\text{att}}}(y_\ell^{(i)} \mid x_\ell^{(i)}) \right], \tag{9}$$

where N is the batch size, and $\left(x_\ell^{(i)}, y_\ell^{(i)}\right)$ are the target language input-output pairs in the batch.

$\mathcal{L}_{\text{CL}}$ is our proposed cross-lingual loss that encourages the model to generate outputs in the correct language based on the input language by utilizing cross-lingual data pairs within the same batch. It is defined as:

$$\begin{aligned} \mathcal{L}_{\text{CL}} = -\frac{1}{N} \sum_{i=1}^{N} \Big[& \log \sigma \left(z_{\text{en}}^{(i)} \right) \\ & + \log \sigma \left(z_\ell^{(i)} \right) \Big], \end{aligned} \tag{10}$$

where $z_{\text{en}}^{(i)}$ and $z_\ell^{(i)}$ are defined respectively as:

$$z_{\text{en}}^{(i)} = \beta \left(\log \frac{\pi_{\theta_{\text{att}}}(y_{\text{en}}^{(i)} \mid x_{\text{en}}^{(i)})}{\pi_{\text{ref}}(y_{\text{en}}^{(i)} \mid x_{\text{en}}^{(i)})} - \log \frac{\pi_{\theta_{\text{att}}}(y_\ell^{(i)} \mid x_{\text{en}}^{(i)})}{\pi_{\text{ref}}(y_\ell^{(i)} \mid x_{\text{en}}^{(i)})} \right), \tag{11}$$

$$z_\ell^{(i)} = \beta \left(\log \frac{\pi_{\theta_{\text{att}}}(y_\ell^{(i)} \mid x_\ell^{(i)})}{\pi_{\text{ref}}(y_\ell^{(i)} \mid x_\ell^{(i)})} - \log \frac{\pi_{\theta_{\text{att}}}(y_{\text{en}}^{(i)} \mid x_\ell^{(i)})}{\pi_{\text{ref}}(y_{\text{en}}^{(i)} \mid x_\ell^{(i)})} \right). \tag{12}$$

In these equations, π_θ denotes the parameterized policy, π_{ref} represents the reference policy, β is a hyperparameter indicating the strength of the KL constraint, σ is the sigmoid function, and the superscript (i) refers to the i-th sample in the batch.

By incorporating both the preferred and rejected responses in different languages within the same batch, the model is explicitly guided to increase the likelihood of outputs in the appropriate language for a given input language while decreasing the likelihood of outputs in the incorrect language. This mechanism ensures that the model not only learns the correspondence between input and output languages but also effectively utilizes its English knowledge for the target language without losing its proficiency in English.

Comparison with Standard DPO on Cross-Lingual Data It's important to emphasize the key differences between our proposed CLO and simply applying the standard DPO to cross-lingual augmented data. In the standard DPO approach with cross-lingual data augmentation, the loss function is applied independently to each language's data, and the model does not utilize cross-lingual data pairs within the same batch. That is, the model may only learn the correspondence between input and output languages *implicitly* and might not effectively utilize the relationships between accepted and rejected responses across languages.

In contrast, our CLO method constructs the loss function by leveraging English and target language data pairs within the *same batch*, as shown in Equations (9)–(12). By pairing each English input-output pair with its corresponding translated target language input-output pair within the batch, the model is explicitly taught to respond in English when given English input and in the target language when given target language input. Moreover, by contrasting the probabilities of the accepted and rejected responses across languages within each sample in the batch, the model prevents knowledge loss and encourages the utilization of English knowledge in the target language.

Through this approach, CLO enables the model to select the correct output language according to the input language and allows for effective target language transfer using its English knowledge. This is fundamentally different from just applying DPO on cross-lingual data augmentation, where the model might not sufficiently learn to adjust the output language based on the input language, potentially leading to suboptimal cross-lingual transfer and loss of English proficiency.

By combining cross-lingual data augmentation with our newly designed batch-based loss function, CLO ensures that the model preserves its English knowledge while effectively transferring it to the target language, achieving superior performance compared to methods that only use data augmentation with standard DPO.

B Hyperparameters

In our experiments, we utilized a fixed setup with a server equipped with 8 NVIDIA A100 GPUs, each with 80GB of memory. The training hyperparameters were set as follows:

- **Trade-off Parameter** (λ): `0.5`
- **Learning rate**: `5e-5`
- **Minimum learning rate**: `1.1e-6`
- **Max sequence length**: `3000`
- **Beta** (β): `0.1`
- **Training batch size**: `8`

All models used in the experiments (except for the SFT + DPO in Table 3) were trained for only 1 epoch. During training, the model with the lowest validation loss was selected for use. The SFT + DPO models experienced an additional training process using the same cross-lingual dataset as CLO during the DPO phase for each SFT model.

For text generation, we used the following fixed generation configurations:

- **Top-p (nucleus sampling)**: `0.9`
- **Temperature**: `0.6`

We set the trade-off parameter $\lambda = 0.5$ for all models and languages in our experiments. While tuning λ for each model and language individually may yield further performance improvements, we use a fixed value of 0.5 in order to report consistent results across different models and languages and to ensure fair comparison.

C AlpacaEval Setup

We detail the prompts used for evaluating and ranking LLMs. To more accurately measure the multilingual capabilities of the models, except for English, we modified the existing prompts for the other languages (Chinese, Korean, and Swahili). Specifically, we included additional instructions to evaluate responses that are in a different language from the instruction. If a model responds in a language different from the one used in the instruction, this is reflected in its evaluation. This adjustment allows us to assess the models' performance in multilingual settings more effectively. Conversely, when evaluating English performance, we used the original AlpacaEval (Li et al., 2023) prompts without modification.

The following is the prompt provided to the `GPT-4o` for evaluation:

System Prompt:

```
You are a highly efficient assistant, who evaluates and rank large language
     models (LLMs) based on the quality of their responses to given prompts.
     This process will create a leaderboard reflecting the most accurate and
     human-preferred answers.
```

User Prompt:

```
I require a leaderboard for various large language models. I'll provide you
     with prompts given to these models and their corresponding responses.
    Your task is to assess these responses, ranking the models in order of
    preference from a human perspective. Once ranked, please output the
    results in a structured JSON format for the make_partial_leaderboard
    function.

## Prompt
{
    "instruction": """{instruction}""",
}

## Model Outputs

Here are the unordered outputs from the models. Each output is associated
    with a specific model, identified by a unique model identifier.
{
    {
        "model": "m",
        "output": """{output_1}"""
    },
    {
        "model": "M",
        "output": """{output_2}"""
    }
}

## Task
Evaluate and rank the models based on the quality and relevance of their
    outputs. The ranking should be such that the model with the highest
    quality output is ranked first. Additionally, since the purpose is to
    measure multilinguality, if a model responds in a language different
    from the instruction's language, this should be reflected in the
    evaluation.
```

The key modification in this prompt is the inclusion of a specific instruction to consider the language of the model's response relative to the instruction's language. By doing so, we aim to evaluate the models' multilinguality more clearly. If a model provides an answer in a language different from the one used in the instruction, this difference is factored into its evaluation, potentially affecting its ranking on the leaderboard.

By employing OpenAI's `GPT-4o` model for the evaluation, we utilize its advanced understanding and reasoning capabilities to perform a nuanced assessment of the LLMs' responses.

D MMMLU Evaluation Setup

In the original MMMLU test set, all fixed prompts requiring answers are written in English. We found that consistently using English prompts without employing language-specific prompts corresponding to the test set resulted in high standard deviations in performance. Therefore, to more accurately measure adaptability and performance in the target language, we adjusted the prompts to match the input language. The specific prompts for each language are presented in Table 6. Each prompt means exactly same instructions, requiring the answers to multiple-choice questions to be written in a specific format by selecting one of A, B, C, or D. The dataset comprises approximately 14,000 data points.

If the correct answer cannot be extracted from the model's response, it is considered incorrect, as this indicates that the model has not properly followed the instructions. (Since MMMLU's performance focuses on evaluating the model's knowledge, the answer is considered correct as long as the correct answer is inferred, even if the given query is not in English and the response is in English.)

We were concerned that, due to limitations in the capabilities of existing pre-trained models, significant performance disparities between models could arise if the correct answer was inferred but could not be extracted from the response. Since extraction might fail due to issues with regular expression matching, we aimed to enhance the reliability of the extraction process. To this end, the evaluation model included 10 English and 10 target language test samples per target language in its training. The 10 test samples used for training were excluded from the MMMLU performance evaluation. If the answer could not be extracted from the multiple-choice responses, we allowed up to three regeneration attempts. If the answer still could not be extracted after these attempts, the response was considered incorrect.

Language	**Prompt**
English	Answer the following multiple choice question. Ensure the last line of your response is in the format: 'Answer: $LETTER' (without quotes) where LETTER is one of ABCD. For example, 'Answer: A'.
Korean	다음 선택형 질문에 답하십시오. 당신의 응답 마지막 줄을 '답변: $LETTER' (따옴표 제외) 형식으로 작성하십시오. 여기서 LETTER는 ABCD 중 하나입니다. 예를 들어, '답변: A'로 대답하세요.
Chinese	回答以下。保的回答的最后一行格式：'答案: $LETTER'（不引), 其中 LETTER 是 ABCD 之一。例如 , '答案: A'。
Swahili	Jibu swali lifuatalo la chaguo nyingi. Hakikisha mstari wa mwisho wa jibu lako uko katika muundo: 'Answer: $LETTER' au 'Jibu: $LETTER' (bila nukuu) ambapo LETTER ni moja ya ABCD. Kwa mfano, 'Answer: A' au 'Jibu: A'.

Table 6: Language-specific prompts for MMMLU test set

E Machine Reading Comprehension Performances

To evaluate the Machine Reading Comprehension (MRC) capabilities of SFT and CLO, we adopt the BELEBELE dataset (Bandarkar et al., 2023). In this benchmark, a question and a passage are provided along with four answer choices. The model's accuracy is determined by calculating the log likelihood of each of the four options and selecting the highest-likelihood answer (Gao et al., 2024).

We describe our language-specific prompt templates designed for our multilingual machine reading comprehension experiments. In order to maintain both a consistent structure across languages and the native labeling of input fields, we map the input fields (passage, question, and answer options) to their corresponding local terms. Table 7 provides a summary of the mappings for each target language along with the respective answer indicator. In our experiments, these templates are used to format the input as a single text string that is passed to the model.

F Analysis of Generation Prompt Strategy

In our main performance evaluation on AlpacaEval (Table 1), the models generated responses without including system prompts instructing them to answer in the respective target languages. According to the study by Marchisio et al. (2024), adding instructions that direct the model to respond in the target

Language	Passage	Question	Option Labels	Answer Indicator
Chinese	文档	问题	A, B, C, D	答案
German	Dokument	Frage	A, B, C, D	Antwort
Korean	문서	질문	가), 나), 다), 라)	정답
Indonesian	Dokumen	Pertanyaan	A, B, C, D	Jawaban
Swahili	Hati	Swali	A, B, C, D	Jibu
Yoruba	Iwe	Ibéèrè	A, B, C, D	Idahun

Table 7: Mapping of Input Fields and Answer Indicators for Each Language in the BELEBELE benchmark.

language can alleviate language confusion. Therefore, we conducted additional experiments to assess the impact of such zero-shot prompts on the models' performance.

Specifically, we included an instruction in the system prompt, such as "Please answer in the same language as the input," translated into the target language. This was intended to encourage the model to generate responses in the appropriate language. We then performed AlpacaEval using these adjusted prompts, and the results are presented in Table 8.

As a result, we observed that even with the inclusion of zero-shot prompts instructing the models to respond in the target language, the overall trends remained similar to those without such prompts. While there were slight improvements in win rates for some models and languages, the performance gains were not substantial. This outcome highlights the limitations of zero-shot prompts in significantly enhancing the target language generation capabilities of the models.

These findings suggest that simply instructing models to answer in the same language as the input is insufficient for overcoming language generation challenges in multilingual contexts.

Model	Eval Language	Chinese	Korean	Swahili
Llama-3-8B	Target	86.1	76.5	70.8
Llama-2-7B	Target	64.7	54.5	64.1
Llama-2-13B	Target	62.0	53.4	74.4
Mistral-7B-v0.1	Target	57.9	50.5	51.3

Table 8: Win rates (%) of **CLO over SFT** on AlpacaEval when models are prompted to answer in the same language as the input using zero-shot prompts.

G Analysis of AlpacaEval Prompt Strategies

In our experiments using AlpacaEval (as shown in Table 1), a comparative analysis between the language-specific prompts used for the target language and the AlpacaEval prompts, which show high correlation with human evaluators, is presented in Table 9.[8]

Apart from two cases in the generated outputs of the Llama-2-7B Chinese model, both prompt types resulted in identical performance, indicating that the language-specific prompts we used also demonstrate a high correlation with human evaluations. Additionally, a manual review revealed that discrepancies in rankings between the evaluations using language-specific prompts and the original AlpacaEval prompts were due to the presence of English entities mixed in the responses, which, when translated or rewritten in the target language, were deemed to be of higher quality. This finding aligns with our intentions, and thus, our main performance results are reported using the AlpacaEval evaluations with language-specific prompts.

[8]Since the original AlpacaEval prompt was used for evaluating English performance, we focus solely on the evaluation of the target language using language-specific prompts.

Model	Language	Language-specific	Original
Llama-2-7B	Chinese	61.1	61.2
	Korean	53.8	53.8
	Swahili	65.0	65.0
Llama-2-13B	Chinese	65.2	65.2
	Korean	53.9	53.9
	Swahili	70.9	70.9
Llama-3-8B	Chinese	83.0	83.0
	Korean	77.8	77.8
	Swahili	70.4	70.4
Mistral-7B-v0.1	Chinese	57.4	57.4
	Korean	50.5	51.5
	Swahili	51.3	51.3

Table 9: Comparison of performance between our language-specific prompts and the original prompts. The table shows the win rate (%) compared to the models trained with SFT by language for each model based on the type of prompt used.

H Analysis of MMMLU Performance in Target Languages

We present a comprehensive analysis of the MMMLU performance of our proposed CLO Llama-3 and Llama-2 7B model compared to the SFT Llama-3 and Llama-2 7B model across various subjects in Chinese, Korean, and Swahili, as illustrated in Figure 4. To conduct a more detailed analysis, we refined the existing 57 categories defined in MMMLU into 24 more specific categories.

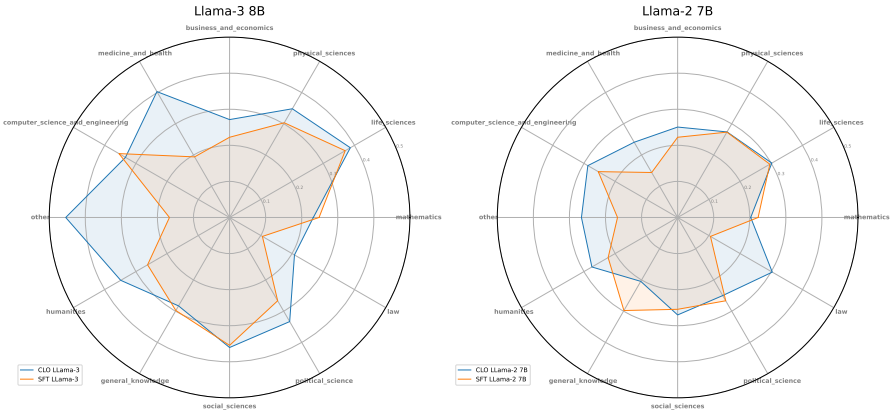

Figure 4: Comparison of average MMMLU performance by category for CLO and SFT models of Llama-2 and Llama-3 in Chinese, Korean, and Swahili languages.

I Effect of Training Data Size on Llama-3

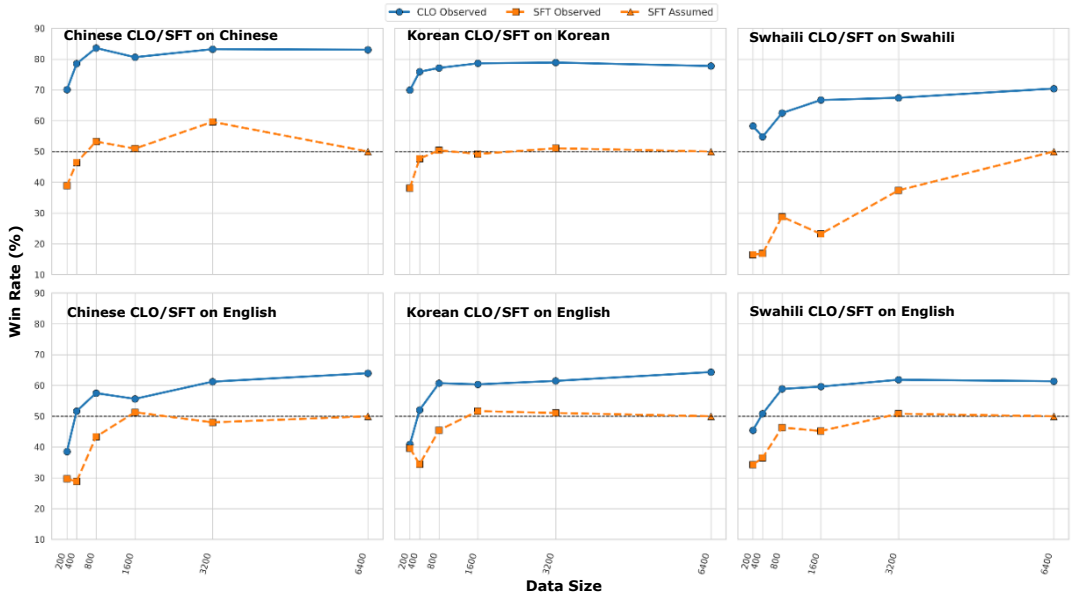

Figure 5: Comparison of win rates between CLO and SFT Llama-3 models, trained with varying amounts of data, against a model fine-tuned using the SFT method with 6,400 pair examples on the AlpacaEval dataset. The "SFT Assumed" baseline is assigned a win rate of 50% since it compares the same model and represents the ideal performance of an SFT model trained with fewer than 6,400 pairs.

J Generation Result Comparison

Query	**1920년대의 멋진 음악에는 어떤 것이 있나요?** (*What is some cool music from the 1920s?*)
Llama2 Chat	*Note: The model responded in English to the Korean query, indicating a language alignment issue.* Thank you for your question! The 1920s were a vibrant and dynamic time for music, with many different styles and genres emerging. Some popular forms of music during this era include: 1. Jazz: This genre originated in the African-American communities of the southern United States and became popular throughout the 1920s. Jazz is characterized by its improvisational nature, syncopated rhythms, and blues and swing influences. 2. Blues: The blues style originated in the Deep South and was popularized by musicians such as Bessie Smith and Ma Rainey. Blues is a genre of music that is characterized by its soulful, emotive sound and 4/4 rhythm. 3. Swing: Swing music emerged in the 1920s and became a popular form of dance music. It is characterized by its strong, syncopated rhythms and lively, energetic sound. 4. Big Band: Big band music was popular in the 1920s and 1930s, and was characterized by its large ensembles and complex arrangements. Big bands typically featured brass, woodwinds, and percussion instruments. 5. Classical: Classical music was also popular in the 1920s, with many composers such as George Gershwin and Leonard Bernstein creating works that blended classical and popular music styles. Some notable musicians and bands from the 1920s include: - Louis Armstrong - Duke Ellington - Bessie Smith - Ma Rainey - Count Basie - Benny Goodman - Artie Shaw - Glenn Miller I hope this information is helpful! Let me know if you have any other questions.
Llama2 SFT	*Note: The model's output is incoherent and appears to be truncated prematurely, resulting in incomplete and incorrect information.* 1. 스 dz 밴드: 1920년대는 음악의 혁신적인 기술적인 사운드와 강렬한 리드 보컬로 유명해진 밴드는 혁신적인 음악을 선보였습니다. 주요 구성원인 스티브 존슨, 루이스 보머, 조니 그린의 음악은 사운드의 원점이 되었습니다. 2. 블루스: 블루스는 아프리카계 미국인의 음악으로, 1920년대에는 주요한 음악 장르가 되었습니다. 주요 블루스 아티스트에는 해리 핀, 로버트 존슨, 차리 매그루드 등이 국제성: 미국 음악의 국제적인 확장.
Llama2 CLO	*Note: The model provides a coherent and contextually appropriate response in Korean.* 1920년대의 멋진 음악에는 다음과 같은 것들이 있습니다: 1. 재즈 음악 - 이 시기는 재즈 음악이 크게 발전한 시기로, 루이 암스트롱, 듀크 엘링턴, 베시 스미스와 같은 아티스트들이 활동했습니다. 이들은 재즈의 황금기를 이끌며 많은 인기를 얻었습니다. 2. 블루스 - 블루스 음악도 이 시기에 큰 발전을 이루었으며, 로버트 존슨과 같은 블루스 뮤지션들이 활동하면서 블루스 음악의 기반을 다졌습니다. 3. 빅 밴드와 스윙 음악 - 1920년대 후반에는 빅 밴드와 스윙 음악이 인기를 끌기 시작했습니다. 베니 굿맨, 카운트 베이시 등의 밴드 리더들이 대형 오케스트라를 이끌며 대중적인 성공을 거두었습니다. 4. 브로드웨이 뮤지컬 - 1920년대는 또한 브로드웨이 뮤지컬의 황금기로, 조지 거슈윈, 콜 포터 등의 작곡가들이 활동하면서 많은 히트곡들을 만들어냈습니다.

Table 10: Comparison of outputs generated by the proposed CLO method, the existing SFT model, and the publicly available Llama-2-7B Chat model in response to the Korean query *"1920년대의 멋진 음악에는 어떤 것이 있나요?"* ("What is some cool music from the 1920s?").