Contrastive Preference Optimization: Pushing the Boundaries of LLM Performance in Machine Translation

Haoran Xu[♠] Amr Sharaf[♡] Yunmo Chen[♠] Weiting Tan[♠] Lingfeng Shen[♠] Benjamin Van Durme[♠]
Kenton Murray[♠ *] Young Jin Kim[♡ *]

Abstract

Moderate-sized large language models (LLMs) – those with 7B or 13B parameters – exhibit promising machine translation (MT) performance. However, even the top-performing 13B LLM-based translation models, like ALMA (Xu et al., 2023), does not match the performance of state-of-the-art conventional encoder-decoder translation models or larger-scale LLMs such as GPT-4 (OpenAI, 2023). In this study, we bridge this performance gap. We first assess the shortcomings of supervised fine-tuning for LLMs in the MT task, emphasizing the quality issues present in the reference data, despite being human-generated. Then, in contrast to supervised fine-tuning which mimics reference translations, we introduce **Contrastive Preference Optimization (CPO)**, a novel approach that trains models to avoid generating adequate but not perfect translations. Applying CPO to ALMA models with only 22K parallel sentences and 0.1% parameters yields significant improvements. The resulting model, called **ALMA-R**, can match or exceed the performance of the WMT competition winners and GPT-4 on WMT'21, WMT'22 and WMT'23 test datasets.[1]

1. Introduction

Machine translation (MT) predominantly utilizes transformer encoder-decoder architectures (Vaswani et al., 2017), which is evident in prominent models such as NLLB-200 (NLLB TEAM et al., 2022), M2M100 (Fan et al., 2021), BIBERT (Xu et al., 2021), and MT5 (Xue et al., 2021). However, the emergence of decoder-only large language models (LLMs) such as the GPT series (Brown et al., 2020; OpenAI, 2023), Mistral (Jiang et al., 2023), LLaMA (Touvron et al., 2023a;b), Falcon (Almazrouei et al., 2023), and others, which have shown remarkable efficacy in various NLP tasks, which attracts the interest of developing machine translation with these decoder-only LLMs. Recent studies (Zhu et al., 2023a; Jiao et al., 2023b; Hendy et al., 2023; Kocmi et al., 2023; Freitag et al., 2023) indicate that larger LLMs such as GPT-3.5 (175B) and GPT-4 exhibit strong translation abilities. However, the performance of smaller-sized LLMs (7B or 13B) still falls short when compared to conventional translation models (Zhu et al., 2023a).

Therefore, there are studies intend to enhance the translation performance for these smaller LLMs (Yang et al., 2023; Zeng et al., 2023; Chen et al., 2023; Zhu et al., 2023b; Li et al., 2023; Jiao et al., 2023a; Zhang et al., 2023), but their improvements are relatively modest, primarily due to the predominant pre-training of LLMs on English-centric datasets, resulting in limited linguistic diversity (Xu et al., 2023). Addressing this limitation, Xu et al. (2023) introduce ALMA models, which initially fine-tune LLaMA-2 (Touvron et al., 2023b) with extensive monolingual data in various languages to enhance their multilingual abilities, and then perform supervised fine-tune (SFT) with small but high-quality parallel data to induce the model toward the translation generation. This method has allowed ALMA models to outperform all prior moderated-size LLMs, and even larger models such as GPT-3.5, in the translation task. Nonetheless, the performance still lags behind leading translation models such as GPT-4 and WMT competition winners. Our study bridges this gap by further fine-tuning ALMA models with our novel training method **Contrastive Preference Optimization (CPO)** and minimal costs, i.e., only 12M learnable parameters (equivalent to 0.1% of the original model size) and a 22K dataset for 10 directions. The fine-tuned model is referred to as **ALMA-R**. A detailed performance comparison is illustrated in Figure 1.

CPO aims to mitigate two fundamental shortcomings of SFT. First, SFT's methodology of minimizing the discrepancy between predicted outputs and gold-standard references inherently caps model performance at the quality level of the

*Equal contribution . Work done during an internship at Microsoft. ♠Johns Hopkins University ♡Microsoft. Correspondence to: Haoran Xu <hxu64@jhu.edu>, Kenton Murray <kenton@jhu.edu>, Young Jin Kim <youki@microsoft.com>.

[1] We release our code and models at: https://github.com/fe1ixxu/ALMA.

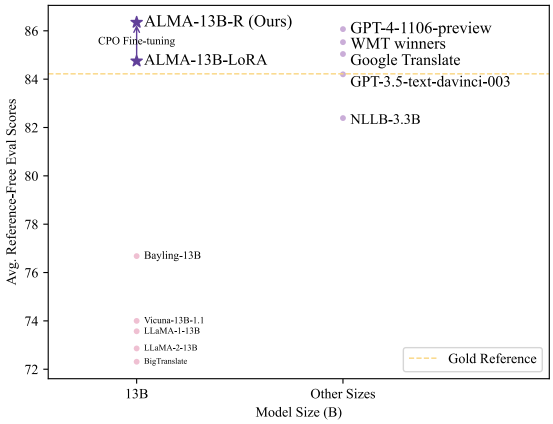

Figure 1: A performance comparison featuring our proposed model ALMA-13B-R against other recently released 13B LLM-based models, as well as top-performing translation systems like GPT-4, WMT winners, Google Translate, and NLLB-200. This evaluation covers the WMT'22 test data across 8 directions, involving translations to and from English for German, Czech, Chinese, and Russian. Scores are averaged based on assessments from three reference-free models: `wmt23-cometkiwi-da-xxl`, `XCOMET-XXL`, and `wmt22-cometkiwi-da`, and are also averaged across all directions. The gold reference is also evaluated due to the reference-free approach. Our model, ALMA-13B-R, developed by further training ALMA-13B-LoRA using our proposed CPO method, either matches or surpasses the most advanced translation models, We show the detailed numerical data for all systems presented in the figure in Appendix A.

training data. This limitation is significant, as even human-written data, traditionally considered high-quality, is not immune to quality issues (more details in Section 2). For instance, one may notice that some strong translation models are capable of producing translations superior to the gold reference, as illustrated in Figure 1. Secondly, SFT lacks a mechanism to prevent the model from rejecting mistakes in translations. While strong translation models can produce high-quality translations, they occasionally exhibit minor errors, such as omitting parts of the translation. *Preventing the production of these near-perfect but ultimately flawed translation is essential*. To overcome these issues, we introduce Contrastive Preference Optimization (CPO) to train the ALMA model using specially curated preference data. After CPO training, the ALMA-R model shows marked improvements, achieving performance levels that match or even surpass those of GPT-4 and WMT competition winners.

Our main contributions are summarized as follows:

Are reference Gold or Gilded? We conducted an in-depth analysis of the training data (FLORES-200 data) utilized by the ALMA model. We meticulously compared the quality of the reference translations with those generated by strong translation models. Our findings reveal that, in numerous instances, the quality of human-written parallel data is even inferior to that of system-generated translations. This observation underscores a critical insight: training models exclusively towards replicating reference translations may not be the most effective approach, and reliance on reference-based evaluation could be flawed.

Pushing the Performance Boundary of SFT We introduce Contrastive Preference Optimization, which offers advantages in terms of memory efficiency, speed, and, crucially, enhanced effectiveness in improving translation quality. CPO breaks the performance bottleneck inherent in SFT's reference-mimicking learning process and push the performance boundary of models that have reached saturation through SFT training.

2. Gold or Gilded? Scrutinizing Gold Reference Quality

The significance of target references is paramount in machine translation tasks. The paradigm of training models on the machine translation task heavily relies on the quality of the references since the model is commonly optimized using a loss that is defined to minimize the difference between the predicted outputs and gold reference. Consider a dataset $\mathcal{D}$, comprising pairs of source sentences x and their corresponding target sentences (gold references) y, represented as $\mathcal{D} = \left\{x^{(i)}, y^{(i)}\right\}_{i=1}^{N}$, where N is the total number of parallel sentences. The negative log-likelihood loss for these parallel sentences, in relation to a model π_θ parameterized by θ, is defined as follows:

$$\mathcal{L}_{\text{NLL}} = -\mathbb{E}_{(x,y)\sim\mathcal{D}}[\log \pi_\theta(y|x)]. \tag{1}$$

Hence, the ability of models to effectively translate is contingent upon the availability of high-quality translation pairs (Xu et al., 2023; Maillard et al., 2023). Furthermore, prevalent evaluation tools such as BLEU (Papineni et al., 2002) and COMET-22 (Rei et al., 2022) predominantly rely on reference-based metrics. However, the precision of these evaluations is sensitive to and compromised by substandard references (Kocmi et al., 2023; Freitag et al., 2023). Recent research (Xu et al., 2023; Kocmi et al., 2023; Freitag et al., 2023) has shifted attention towards assessing the quality of parallel datasets, indicating that target references may not consistently represent the highest quality. In Figure 2, we take a translation example from the FLORES-200 dataset, and compare the gold reference translation with outputs from the best ALMA model and GPT-4. This comparison reveals that the gold reference is a flawed translation, as it omits part of information, whereas the system-generated outputs demonstrate superior quality. This prompts an inquiry: *Are references (even though human-written) truly equivalent to gold standards?* To thoroughly assess the quality of both the gold standard references and the outputs from contemporary high-performance translation models, we propose evaluating these outputs utilizing reference-free evaluation frameworks.

Source: 这是马特利 (Martelly) 四年来第五次入选海地临时选举委员会 (CEP)。

Reference: It is Martelly's fifth CEP in four years.

ALMA-13B-LoRA: This is Martelly's fifth time **being selected by the Provisional Electoral Council** (CEP) in four years.

GPT-4: This is the fifth time Martelly has been **selected for Haiti's Provisional Electoral Council** (CEP) in four years.

Figure 2: An example demonstrating that a human-written gold reference may not always be flawless, and could be surpassed by translations from advanced translation models. In this case, the reference retains the abbreviation "CEP" but fails to provide its full name. The highlighted phrases in the model-generated translations indicate the portions omitted by the gold reference.

Models We scrutinize the translation outputs from ALMA-13B-LoRA[2], as well as zero-shot translations from the most recent GPT-4 (`gpt-4-1106-preview`). To assess the quality of these outputs, we employ two of the latest and largest reference-free models, each with a 10B parameter size and demonstrating very high correlation with human judgements (Freitag et al., 2023). These models are `Unbabel/wmt23-cometkiwi-da-xxl` (henceforth referred to as **KIWI-XXL**) (Rei et al., 2023) and `Unbabel/XCOMET-XXL` (subsequently referred to as **XCOMET**) (Guerreiro et al., 2023).

Data we consider the high-quality and human-written FLORES-200 dataset (NLLB TEAM et al., 2022), comprising both development and test data, amounting to a total of 2009 samples for each language direction, to compare the gold references with the outputs generated by the models. We employed ALMA-13B-LoRA and GPT-4 to perform translations across five English-centric language pairs, covering both translations from and to English. These pairs include German (`de`), Czech (`cs`), Icelandic (`is`), Chinese (`zh`), and Russian (`ru`), with Icelandic (`is`) categorized as a low-resource language and the others as high-resource languages.

Prompt The prompt employed for generating translations with ALMA models is consistent with the one used in Xu et al. (2023). For GPT-4 translation generation, we follow the guidelines suggested by Hendy et al. (2023). The specifics of these prompts are detailed in Appendix B.

[2]ALMA-13B-LoRA is the best 13B translation model in the ALMA families. It initially undergoes *full-weight* fine-tuning on monolingual data, followed by fine-tuning on high-quality human-written parallel data using *low-rank adaptation* (LoRA) (Hu et al., 2022).

Table 1: A performance comparison between gold references and outputs from advanced translation models, as assessed by two 10B-size reference-free evaluation models with the highest correlation to human preferences. The results indicate that the average performance of these strong translation models can even exceed that of the gold references, achieving a high success rate in beating the reference.

	KIWI-XXL	Win Ratio (%)	XCOMET	Win Ratio (%)
Translating to English (xx→en)				
Reference	85.31	-	88.82	-
ALMA-13B-LoRA	88.33	73.24	92.68	60.17
GPT-4	89.21	79.43	94.66	54.25
Translating from English (en→xx)				
Reference	87.85	-	94.42	-
ALMA-13B-LoRA	85.62	42.15	93.07	35.46
GPT-4	87.30	49.13	94.21	38.09

Model Outputs Can Be Better References In Table 1, we present the evaluation scores of KIWI-XXL and XCOMET for the gold references, ALMA-13B-LoRA outputs, and GPT-4 outputs. Additionally, we report *Win Ratio*, reflecting the proportion of instances where model outputs surpass the gold standard references. These metrics are calculated as an average across five languages. Remarkably, even comparing with the high-quality Flores-200 dataset, the average performance of translation models in `xx`→`en` translations significantly exceeds that of the references, showing approximately 3-4 point increases in KIWI-XXL and 4-6 point gains in XCOMET. Notably, a significant proportion of outputs are rated higher than the references by KIWI-XXL (e.g., **73.24%** for ALMA), with a slightly reduced yet still substantial percentage when assessed using XCOMET (**60.17%** for ALMA). In the `en`→`xx` direction, while the overall performance between the translations from reference and two systems is comparable, approximately 40% are still deemed superior to the reference translations.

Motivation: Help The Model Learn Rejection The aforementioned findings illustrate that translations produced by advanced models can sometimes surpass the quality of gold standard references. This raises the question of how to effectively utilize such data. A straightforward approach would involve fine-tuning the model using the source and the superior translations as references. While this could enhance the model's translation abilities, it does not equip the model with the discernment to identify and avoid generating suboptimal translations, exemplified by the "good but not perfect" translations depicted in Figure 2. Consequently, this situation motivates us to develop a new training objective, which aims to instruct the model in prioritizing the generation of higher-quality translations and rejecting lesser ones, in a style of contrastive learning with hard negative examples (Oord et al., 2018; Chen et al., 2020; He et al., 2020; Robinson et al., 2021; Tan et al., 2023). This objective moves beyond the traditional focus on merely minimizing cross-entropy loss towards the reference.

3. Contrastive Preference Optimization

In this section, we present a novel preference learning technique, termed **Contrastive Preference Optimization (CPO)**. This method is designed to guide the model in developing a propensity for generating 'better' translations while simultaneously learning to reject 'worse' ones, even in cases where these 'worse' translations are of high quality but not perfect.

3.1. Deriving the CPO Objective

We discuss the derivation of CPO objective, beginning with an analysis of Direct Preference Optimization (DPO) (Rafailov et al., 2023). DPO represents a more direct optimization objective utilized in reinforcement learning from human feedback (RLHF) (Ziegler et al., 2019; Ouyang et al., 2022). Given a set of source sentences x, alongside preferred translation targets y_w and less preferred ones y_l, we can access a static dataset of comparisons, denoted as $\mathcal{D} = \left\{x^{(i)}, y_w^{(i)}, y_l^{(i)}\right\}_{i=1}^{N}$. The loss function for DPO is constructed as a maximum likelihood objective for a parameterized policy π_θ:

simple enough:
stay away from wrong one
stay close with the right one
— problem: right one is not right enough

$$\mathcal{L}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\sigma\Big(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\Big)\Big], \tag{2}$$

where π_{ref} is a pre-trained language (translation) model , σ is the Sigmoid function, and β is a hyperparameter. The DPO loss is derived a reparameterization process of the ground-truth reward and the corresponding optimal policy in the Proximal Policy Optimization (PPO) framework (Schulman et al., 2017). As a result, DPO training can be conducted in a supervised fine-tuning style, as it relies exclusively on labeled preference data and does not require interaction between agents and their environment.

However, DPO has notable drawbacks compared to common SFT. Firstly, DPO is **memory-inefficient**: it necessitates twice the memory capacity to simultaneously store both the parameterized policy and the reference policy. Secondly, it is **speed-inefficient**: executing the model sequentially for two policies doubles the processing time. To address these inefficiencies, we introduce contrastive preference optimization.

Initially, we set $\pi_{\text{ref}} = \pi_w$, representing an ideal policy that perfectly aligns the true data distribution of the preferred data. Specifically, for any given data point (x, y_w, y_l) from the dataset $\mathcal{D}$, the conditions $\pi_w(y_w|x) = 1$ and $0 \leq \pi_w(y_l|x) \leq 1$ hold true. This contrasts with the conventional approach of assigning π_{ref} to the initial pre-trained model checkpoint. This modification is feasible because the primary role of π_{ref} is to prevent deviation from the original model, and our approach similarly aligns with this goal, maintaining the model's proximity to the ideal preferred model. Consequently, under this setup, the predictions for preferred data do not require reweighting by the reference model, and Equation 2 can be reformulated as follows :

we just want the winning case to be more likely to happen
but do not really ask it to be 'close' to the reference policy
which provides such potentially-flawing GTdata points?

$$\mathcal{L}(\pi_\theta; \pi_w) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\sigma\Big(\beta\log\pi_\theta(y_w|x) - \beta\log\pi_\theta(y_l|x) + \beta\log\pi_w(y_l|x)\Big)\Big]. \tag{3}$$

However, $\beta\log\pi_w(y_l|x)$ is unknown as the ideal model π_w is unreachable, but we can approximate the optimization. After expanding the Sigmoid function and removing the non-parameterized term, the loss becomes to:

$$\mathcal{L}(\pi_\theta; \pi_w) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\pi_\theta(y_w|x)^\beta - \log\Big(\pi_\theta(y_w|x)^\beta\cdot\pi_w(y_l|x)^\beta + \pi_\theta(y_l|x)^\beta\Big)\Big]. \tag{4}$$

Considering that $0 \leq \pi_w(y_l|x) \leq 1$, the loss can be upper bounded as follows:

$$\mathcal{L}(\pi_\theta; \pi_w) \leq -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\pi_\theta(y_w|x)^\beta - \log\Big(\pi_\theta(y_w|x)^\beta\cdot 1 + \pi_\theta(y_l|x)^\beta\Big)\Big] \tag{5}$$

Upon restructuring the above loss into a Sigmoid function, we derive our new preference learning objective without computing $\beta\log\pi_w(y_l|x)$ by minimizing the upper bound of Equation 4:

$$\mathcal{L}_{\text{prefer}} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\sigma\Big(\beta\log\pi_\theta(y_w|x) - \beta\log\pi_\theta(y_l|x)\Big)\Big]. \tag{6}$$

this is standard contrastive learning target, maximize good case occurance and minimize bad case occurance

A detailed step-by-step derivation from Equation 3 to 6 is provided in Appendix C. Our approach also involves implementing a straightforward signal to guide the learnable policy π_θ towards the preferred data distribution. Specifically, we incorporate a log-likelihood supervised fine-tuning loss applied to the preferred data:

$$\mathcal{L}_{\text{NLL}} = -\mathbb{E}_{(x,y_w)\sim\mathcal{D}}[\log\pi_\theta(y_w|x)]. \tag{7}$$

By combing Equation 6 and 7, our CPO loss is finally formulated as follows:

$$\mathcal{L}_{\text{CPO}} = \mathcal{L}_{\text{prefer}} + \mathcal{L}_{\text{NLL}}. \tag{8}$$

The transition from DPO to CPO resolves issues of memory and speed inefficiency. This efficiency is achieved as CPO only requires the storage and processing of one policy π_θ, so CPO facilitates preference learning with the same space and time complexity as common SFT methods. We also show the substantially superior performance of CPO compared to DPO in Section 4.

It's more about the Data Quality here....

3.2. Triplet Preference Data

We here describe the construction of our preference data $\mathcal{D}$. This dataset is developed using the FLORES-200 data (both development and test sets) and encompasses the same language pairs as discussed in Section 2. For each language pair, the dataset comprises 2009 parallel sentences.

For a given source sentence x, whether translated from or to English, we utilize both GPT-4 and ALMA-13B-LoRA to generate respective translations, denoted as $y_{\text{gpt-4}}$ and y_{alma}. Together with the original target reference y_{ref}, this forms a triplet $\mathbf{y} = (y_{\text{ref}}, y_{\text{gpt-4}}, y_{\text{alma}})$, representing three different translation outputs for the input x. The reference-free evaluation models KIWI-XXL and XCOMET are then employed to score these translations, with the average scores represented as $\mathbf{s} = (s_{\text{ref}}, s_{\text{gpt-4}}, s_{\text{alma}})$.[3] The highest-scoring translation is labeled as the preferred translation y_w, and the lowest-scoring as the dis-preferred translation y_l, i.e., $y_w = \mathbf{y}_{\arg\max_i(\mathbf{s})}, y_l = \mathbf{y}_{\arg\min_i(\mathbf{s})}$, where i represents the index in the triplet. Translations with intermediate scores are not considered. An illustrative example of this selection process is depicted in Figure 3. It is important to note that even the dis-preferred translations may be of high-quality. The designation 'dis-preferred' indicates that there is still room for improvement, perhaps through the addition of minor details. This approach of using high-quality but not flawless translations as dis-preferred data aids in training the model to refine details and achieve perfection in generated translations.

4. Experiments

4.1. Data

Following Section 2, we consider 10 translation directions in the paper: cs↔en, de↔en, is↔en, zh↔en, ru↔en. Building on the ALMA models' (Xu et al., 2023) insights that a small quantity of high-quality data can yield impressive translation results, our training dataset is even more compact. As detailed in Section 3.2, our preference training data is derived from the FLORES-200 dataset, a subset of which has been also employed in the training of ALMA models. This results in a total of 2K × 10 directions = 20K paired sentences. In addition to preference data assessed

dataset quality means everything

[3] The impact of using different evaluation models, such as only using XCOMET or KIWI-XXL, is explored in Section 5.1.

Figure 3: A triplet of translations, either model-generated or derived from a reference, accompanied by their respective scores as assessed by reference-free models. For a given source sentence, the translation with the highest score is designated as the preferred translation, while the one with the lowest score is considered dis-preferred, and the translation with a middle score is disregarded.

by large evaluation models, our dataset incorporates 1K internal human-labeled preference data, containing preferred and dis-preferred translations along with human preference. However, the human-labeled data is limited to just two translation directions: en→zh and en→de. The details regarding the composition and influence of human-labeled data are explored in Appendix D.[4] In alignment with Xu et al. (2023), our primary focus is on the test set drawn from WMT'21 for is and WMT'22 for other languages. Additionally, we conduct auxiliary experiments evaluating models on WMT'23, covering six directions: de↔en, zh↔en, and ru↔en.

4.2. Training Setup

We train the model in a *many-to-many* multilingual machine translation manner, starting with ALMA-13B-LoRA as the initial checkpoint. During the training phase, we focus exclusively on updating the weights of the added LoRA parameters. These weights have a rank of 16 and only add an additional 12M parameters to the original 13B size of the model. We adhere to the default β value of 0.1 as suggested by Rafailov et al. (2023). The fine-tuning process of ALMA-13B-LoRA involves a batch size of 128, a warm-up ratio of 0.01, spanning a single epoch, and accommodating sequences with a maximum length of 512 tokens. To optimize training efficiency, we integrate the deepspeed tool (Rasley et al., 2020). We utilize the same prompt as Xu et al. (2023) and do not compute the loss for the prompt. While our primary focus is on the performance of 13B models, CPO markedly benefits 7B models as well. Consequently, we also release ALMA-7B-R and provide a detailed discussion

[4] TL;DR: A brief overview of the impact of this human-labeled data suggests a minimal effect.

Table 2: The overall results in en→xx for WNT'21 and WMT'22. The application of the CPO method to fine-tune the ALMA-13B-LoRA model leads to a significant enhancement in performance, equalling or surpassing that of WMT competition winners and GPT-4. **Bold** numbers denote the highest scores across all systems. Deep green boxes highlight improvements exceeding 1 point over the ALMA model after fine-tuning with preference data, while more modest gains under 1 point are shown in shallow green boxes. Decreases in performance are marked with red boxes.

Models	de			cs			is		
	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET
Gold Reference	82.67	84.01	**97.85**	83.19	81.83	90.27	80.51	85.20	91.52
WMT Winners	**83.56**	83.70	96.99	85.31	**87.27**	**94.38**	81.77	84.94	91.61
GPT-4	83.48	**84.91**	97.56	84.81	85.35	93.48	81.03	81.21	90.00
ALMA-13B-LoRA	82.62	81.64	96.49	84.14	84.24	92.38	81.71	83.31	91.20
+ SFT on preferred data	82.75	81.85	96.67	84.14	83.46	91.99	81.48	82.11	90.30
+ DPO	82.40	81.20	96.40	83.86	83.45	91.68	81.43	82.66	90.33
+ CPO (Ours, ALMA-13B-R)	83.28	84.25	97.48	**84.99**	87.06	93.61	**82.18**	**85.68**	**91.93**
Models	zh			ru			Avg.		
	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET
Gold Reference	80.92	81.70	90.42	82.96	84.62	94.17	82.05	83.47	92.85
WMT Winners	82.04	81.13	91.14	**84.35**	87.01	94.79	**83.41**	84.81	93.78
GPT-4	81.73	81.53	90.79	83.64	86.15	94.3	82.94	83.83	93.23
ALMA-13B-LoRA	80.82	79.96	89.92	83.10	84.17	93.79	82.48	82.66	92.76
+ SFT on preferred data	81.25	80.51	90.18	83.23	84.15	93.54	82.57	82.42	92.54
+ DPO	80.74	79.64	89.58	82.94	83.40	93.25	82.27	82.07	92.25
+ CPO (Ours, ALMA-13B-R)	**82.25**	**84.32**	**92.03**	83.98	**87.37**	**95.22**	83.34	**85.74**	**94.05**

of its performance in Appendix A.

4.3. Baselines

SoTA Models In this category, our benchmarks are established against, to the best of our knowledge, the strongest publicly available translation models. We first compare with **ALMA-13B-LoRA**, recognized as one of the top moderate-size language-model based translation systems, surpassing notable conventional models such as NLLB-54B in both WMT'21 and WMT'22. We also compare our results with **TowerInstruct**[5], a recently released LLM-based translation model and a contemporary work in the field.[6] Additionally, we evaluate against the zero-shot performance of the latest **GPT-4** (`gpt-4-1106-preview`), currently shown to be the best translation model among all LLM-based translation systems (Xu et al., 2023; Zhang et al., 2023; Zeng et al., 2023; Jiao et al., 2023a). Lastly, we include comparisons with the **WMT competition winners**, representing the highest standard of translation models within the competition, though it is noted that the winning models vary across different language directions.[7]

SFT and DPO We also compare different training objectives. Given that CPO is designed to steer learning towards preferred data, a straightforward benchmark is to compare its performance against directly SFT on the same preferred data set. Furthermore, considering that CPO is an evolution of DPO, we also include a comparative analysis with DPO.

4.4. WMT'21 and WMT'22 Results

We present the primary results for en→xx and xx→en in Table 2 and Table 3, respectively. Our emphasis is primarily on reference-free evaluation models, due to our analysis in Section 2, which questions the reliability of gold references and highlights that evaluations can be compromised by poor-quality references (Kocmi et al., 2023; Freitag et al., 2023). These models include KIWI-XXL, XCOMET, and a smaller yet popular model, `Unbabel/wmt22-cometkiwi-da` (hereinafter referred to as **KIWI-22**). Scores highlighted in **bold** represent the highest achieved across all systems. For a comprehensive comparison, we also include reference-based evaluations using sacreBLEU (Post, 2018) and COMET-22 (`Unbabel/wmt22-comet-da`) (Rei et al., 2022) in Appendix A.

Comparing With SoTA Models While ALMA-13B-LoRA ranks as one of the top moderate-size LLM translation models, it slightly trails behind GPT-4 and the WMT competition winners. However, the incorporation of CPO significantly enhances ALMA's capabilities, bringing its performance to a level that is comparable to or even surpasses that of GPT-4 and WMT winners. For example, ALMA-13B-R achieves an average score of 85.74 on KIWI-XXL and 94.05 on XCOMET for en→xx translations. These scores outperform GPT-4, which scores 83.83 on KIWI-XXL and 93.23 on XCOMET, as well as the WMT winners, who score 84.81

[5] `https://huggingface.co/datasets/Unbabel/TowerBlocks-v0.1`.

[6] Note that TowerInstruct has used WMT'22 test data for training, so we exclude it from comparison on the WMT'22 test dataset.

[7] The WMT winner systems used for comparison in each direction are provided in Appendix E.

Table 3: The overall results in xx→en for WMT'21 and WMT'22. The usage of color and boldface are the same in Table 2.

Models	de			cs			is		
	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET
Gold Reference	78.74	78.56	88.82	82.08	83.11	84.60	80.88	85.04	76.16
WMT Winners	81.38	83.59	93.74	82.47	82.53	85.65	81.39	85.60	78.14
GPT-4	**81.50**	**84.58**	**94.47**	82.52	83.55	**88.48**	81.49	**85.90**	**81.11**
ALMA-13B-LoRA	81.14	83.57	93.30	81.96	82.97	83.95	80.90	85.49	76.68
+ SFT on preferred data	81.36	83.98	93.84	82.36	83.15	86.67	81.32	85.61	80.20
+ DPO	81.13	83.52	93.25	81.82	82.69	83.84	80.89	85.22	76.09
+ CPO (Ours, ALMA-13B-R)	**81.50**	83.97	94.20	**82.63**	**83.75**	88.03	**81.57**	85.73	80.49
Models	zh			ru			Avg.		
	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET
Gold Reference	77.09	74.19	90.70	80.74	79.59	88.56	79.91	80.10	85.77
WMT Winners	77.66	73.28	87.2	81.71	80.97	90.91	80.92	81.19	87.13
GPT-4	**79.33**	**77.65**	**92.06**	81.57	81.34	90.95	81.28	**82.60**	**89.41**
ALMA-13B-LoRA	77.32	74.41	89.88	81.31	81.05	89.89	80.53	81.50	86.74
+ SFT on preferred data	78.32	76.03	90.65	81.46	81.17	90.65	80.96	81.99	88.40
+ DPO	77.50	74.50	89.94	81.19	80.88	89.76	80.51	81.36	86.58
+ CPO (Ours, ALMA-13B-R)	79.24	77.17	91.65	**81.72**	**81.54**	**91.18**	**81.33**	82.43	89.11

Table 4: The average performance in WMT'23 across all six directions, with the highest score among all systems highlighted in bold.

	Avg.		
	KIWI-22	KIWI-XXL	XCOMET
Gold Reference	78.74	75.56	86.30
WMT Winners	**80.57**	77.72	88.24
TowerInstruct	80.31	77.18	88.11
ALMA-13B-LoRA	79.48	76.00	87.16
+ CPO (Ours, ALMA-13B-R)	80.55	**78.97**	**89.74**

on KIWI-XXL and 93.78 on XCOMET.

Comparing With SFT and DPO All training objectives in our study are fine-tuned using the ALMA-13B-LoRA model as a base. In Table 2 and 3, we employ deep green boxes to indicate improvements exceeding 1 point, and shallow green boxes for improvements less than 1 point. Conversely, red boxes denote a decline in performance. We observe that SFT on preferred data marginally enhances the ALMA model's translation capability for xx→en, and results in a slight deterioration for en→xx. Similarly, DPO slightly decreases model performance. In contrast, CPO demonstrates significant improvements across all translation directions.

4.5. WMT'23 Results

We show the average results across all six directions in Table 4, and provide the performance in each direction in Appendix F due to the space constraints. Consistent with observations from WMT'21 and WMT'22, ALMA-13B-R surpasses contemporary moderate-size LLM-based translators such as ALMA-13B-LoRA and TowerInstruct, and either matches or exceeds the performance of WMT winners.

Table 5: The influence of employing various reference-free models for creating preference data. The results illustrates that the final performance disparities are minimal whether using solely KIWI-XXL, XCOMET, or their combined ensemble.

Models for Building Preference Data	KIWI-22	KIWI-XXL	XCOMET
Translating to English (xx→en)			
N/A (ALMA-13B-LoRA baseline)	80.53	81.50	86.74
KIWI-XXL	**81.33**	82.59	88.82
XCOMET	81.27	82.33	**89.17**
Ensemble of above (Original)	**81.33**	**82.43**	89.11
Translating from English (en→xx)			
N/A (ALMA-13B-LoRA baseline)	82.48	82.66	92.76
KIWI-XXL	83.31	85.87	93.97
XCOMET	83.09	85.43	**94.09**
Ensemble of above (Original)	**83.34**	**85.74**	94.05

5. Analyses

All analyses were conducted using the WMT'21 and WMT'22 test datasets, with their averaged performance being reported.

5.1. Are Translations Really Better or Just Metric-Preferred?

In our study, since the preferred data is selected by reference-free models and the same models are used for evaluation, we investigate the potential for 'cheating' in the scoring process. Specifically, we question whether the improved translation scores reflect genuinely better translations or if they simply align more closely with the evaluation model's preferences? This inquiry is addressed in two parts:

At the metric level, we examine if training a model on data preferred by a specific metric (such as KIWI-XXL) yields improvements that are consistent across other metrics. To investigate this, we reconstruct the preference data using

Table 6: An ablation study evaluating the significance of individual components in the CPO loss function, specifically analyzing how the preference learning loss $\mathcal{L}_{\text{prefer}}$ and the log-likelihood loss $\mathcal{L}_{\text{NLL}}$ each contribute to enhancing translation performance.

Loss Objective	KIWI-22	KIWI-XXL	XCOMET
Translating to English (xx→en)			
$\mathcal{L}_{\text{prefer}}$	80.51	81.35	86.58
$\mathcal{L}_{\text{NLL}}$	80.96	81.99	88.40
$\mathcal{L}_{\text{prefer}} + \mathcal{L}_{\text{NLL}}$	**81.33**	**82.43**	**89.11**
Translating from English (en→xx)			
$\mathcal{L}_{\text{prefer}}$	82.23	82.06	92.22
$\mathcal{L}_{\text{NLL}}$	82.57	82.42	92.54
$\mathcal{L}_{\text{prefer}} + \mathcal{L}_{\text{NLL}}$	**83.34**	**85.74**	**94.05**

only KIWI-XXL or XCOMET and re-train the ALMA-13B-LoRA model using the CPO method. The results, presented in Table 5, do not indicate a significant bias towards the metric used for selecting preferred data. We observed similar and consistent improvements across all metrics, regardless of the specific metric used to select the preferred data.

At the method level, we question whether training on metric-preferred data always leads to better scores on that metric, regardless of the method we use. However, the connection is not straightforward; for instance, SFT on preferred data paradoxically results in diminished performance across all three metrics as shown in Table 2.

Consequently, our analysis supports the robustness and validity of using reference-free models like KIWI-XXL and XCOMET both for constructing preference data and for evaluation purposes, underscoring the absence of bias in this approach. Furthermore, Table 5 demonstrates that the choice between using KIWI-XXL, XCOMET, or an ensemble of both has a minimal impact on the results.

5.2. Ablation Study

CPO Loss Components The CPO loss function consists of two components: $\mathcal{L}_{\text{prefer}}$ for preference learning, and $\mathcal{L}_{\text{NLL}}$, which ensures the model does not deviate significantly from the preferred data distribution. To illustrate the significance of each term, we re-train the model exclusively with one of the components. It is important to note that training solely with $\mathcal{L}_{\text{NLL}}$ equates to the baseline scenario of SFT on preferred data. As depicted in Table 6, the inclusion of both terms yields the optimal performance, while the absence of either leads to a decrease in performance.

Preference Data Components: Our preference data selection involves choosing preferred and dis-preferred translations from a triplet consisting of outputs from GPT-4, ALMA, and the gold reference. In Table 7, we emphasize the significance of the data generated by both ALMA and GPT-4. The results indicate a notable decline in performance when ALMA data is excluded, particularly in the en→xx direction, with the KIWI-XXL score dropping from 85.74 to 84.50 on average. Conversely, omitting GPT-4 data leads to a significant performance decrease in the xx→en direction. This demonstrates that data generated by both systems plays a helpful role in enhancing model performance.

Table 7: An ablation study assessing the significance of each component in the translation triplet. By excluding either ALMA or GPT-4 generated data from the preference triplet and re-training the model, we evaluate their respective impacts. The findings highlight the importance of ALMA-generated data for en→xx translations and GPT-4 generated data for xx→en translations.

Preference Data	KIWI-22	KIWI-XXL	XCOMET
Translating to English (xx→en)			
ALMA + Ref	81.10	82.21	87.80
GPT-4 + Ref	81.20	82.28	89.12
All of them	**81.33**	**82.43**	**89.11**
Translating from English (en→xx)			
ALMA + Ref	82.58	85.30	93.09
GPT-4 + Ref	82.37	84.50	93.10
All of them	**83.34**	**85.74**	**94.05**

Table 8: An examination of the impact of dis-preferred data quality, contrasting noised data with natural, high-quality translations receiving the lowest scores as dis-preferred data. The findings underscore the importance of natural and high-quality dis-preferred data.

Dis-Preferred Data	KIWI-22	KIWI-XXL	XCOMET
Translating to English (xx→en)			
Manually Noised	81.01	82.18	88.23
Natural (Ours)	**81.33**	**82.43**	**89.11**
Translating from English (en→xx)			
Manually Noised	82.71	83.13	92.80
Natural (Ours)	**83.34**	**85.74**	**94.05**

5.3. Does The Quality of Dis-preferred Data Matter?

In our experimental setup, dis-preferred data, though originating from strong translation models, receives the lowest scores when compared with two other translation outputs. A pertinent question arises: does the quality of dis-preferred data significantly impact model performance, and can high-quality (albeit imperfect) dis-preferred data aid in translation improvement? To explore this, we constructed a new set of preference data where the dis-preferred translations (y_l) are artificially generated, as opposed to being naturally derived high-quality translations.

In this new dataset, the preferred translation (y_w) remains the best of the three translation candidates, selected in the same manner as in Section 3.2. However, the dis-preferred translation is intentionally modified to be a noised version of y_w. We applied random deletions of words with a probability of 0.15 and word swaps within a range of 1 with a probability of 0.3, following the method suggested by Zeng et al. (2023) for creating manually noised dis-preferred data. This approach produces worse translations that are unnatural and artificial.

Table 8 compares the performance when using these manually noised dis-preferred data versus the original, naturally occurring high-quality dis-preferred data. The results show a substantial decline in performance across all three metrics and both translation directions when the dis-preferred data is manually noised, underscoring the importance of the quality of dis-preferred data in enhancing translation performance.

6. Conclusion

In this study, we initially proposed the potential quality issues of gold references in machine translation tasks, highlighting instances where advanced translation models outperform these references. This finding challenges the conventional assumption of gold references as the optimal standard, impacting not only model training, which often relies on minimizing the difference between predicted tokens and gold references, but also potentially skewing results in reference-based evaluation metrics. Subsequently, we introduce Contrastive Preference Optimization, a more efficient variant of of DPO. This method leverages both model-generated and reference data to guide the model in avoiding near-perfect yet flawed translations and learning superior ones. Our developed model, ALMA-13B-R, stands out as the first moderate-size LLM-based translation model to match, and in some cases surpass, the performance of cutting-edge systems such as GPT-4 and WMT competition winners, marking a significant advancement in the field of neural machine translation.

Acknowledgements

We extend our gratitude to Hieu Hoang, Marcin Junczys-Dowmunt, Huda Khayrallah, Thamme Gowda, Vikas Raunak, Matt Post, Anoop Kunchukuttan, Roman Grundkiewicz, Philipp Koehn, Hany Hassan Awadalla, Arul Menezes and Vishal Chowdhary for their insightful and valuable suggestions.

References

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, E., Heslow, D., Launay, J., Malartic, Q., Noune, B., Pannier, B., and Penedo, G. Falcon-40B: an open large language model with state-of-the-art performance. 2023.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Chen, Y., Liu, Y., Meng, F., Chen, Y., Xu, J., and Zhou, J. Improving translation faithfulness of large language models via augmenting instructions. *arXiv preprint arXiv:2308.12674*, 2023.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., et al. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22 (107):1–48, 2021.

Freitag, M., Mathur, N., Lo, C.-k., Avramidis, E., Rei, R., Thompson, B., Kocmi, T., Blain, F., Deutsch, D., Stewart, C., Zerva, C., Castilho, S., Lavie, A., and Foster, G. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C. (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 578–628, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.51. URL https://aclanthology.org/2023.wmt-1.51.

Guerreiro, N. M., Rei, R., van Stigt, D., Coheur, L., Colombo, P., and Martins, A. F. xcomet: Transparent machine translation evaluation through fine-grained error detection. *arXiv preprint arXiv:2310.10482*, 2023.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., and Awadalla, H. H. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*, 2023.

Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Jiao, W., Huang, J.-t., Wang, W., He, Z., Liang, T., Wang, X., Shi, S., and Tu, Z. ParroT: Translating during chat using large language models tuned with human translation and feedback. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 15009–15020, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.1001. URL https://aclanthology.org/2023.findings-emnlp.1001.

Jiao, W., Wang, W., Huang, J.-t., Wang, X., and Tu, Z. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*, 2023b.

Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., Haddow, B., Koehn, P., Marie, B., Monz, C., Morishita, M., Murray, K., Nagata, M., Nakazawa, T., Popel, M., Popović, M., and Shmatova, M. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C. (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 1–42, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.wmt-1.1.

Li, J., Zhou, H., Huang, S., Chen, S., and Chen, J. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. *arXiv preprint arXiv:2305.15083*, 2023.

Maillard, J., Gao, C., Kalbassi, E., Sadagopan, K. R., Goswami, V., Koehn, P., Fan, A., and Guzman, F. Small data, big impact: Leveraging minimal data for effective machine translation. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2740–2756, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.154. URL https://aclanthology.org/2023.acl-long.154.

NLLB TEAM, Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

OpenAI. Gpt-4 technical report, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In Isabelle, P., Charniak, E., and Lin, D. (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

Post, M. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL https://aclanthology.org/W18-6319.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

Rasley, J., Rajbhandari, S., Ruwase, O., and He, Y. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3505–3506, 2020.

Rei, R., C. de Souza, J. G., Alves, D., Zerva, C., Farinha, A. C., Glushkova, T., Lavie, A., Coheur, L., and Martins, A. F. T. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.52.

Rei, R., Guerreiro, N. M., Pombal, J., van Stigt, D., Treviso, M., Coheur, L., de Souza, J. G., and Martins, A. F. Scaling up cometkiwi: Unbabel-ist 2023 submission for the quality estimation shared task. *arXiv preprint arXiv:2309.11925*, 2023.

Robinson, J. D., Chuang, C.-Y., Sra, S., and Jegelka, S. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=CR1XOQ0UTh-.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Tan, W., Heffernan, K., Schwenk, H., and Koehn, P. Multilingual representation distillation with contrastive learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1469–1482, 2023.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wu, Y. and Hu, G. Exploring prompt engineering with GPT language models for document-level machine translation: Insights and findings. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C. (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 166–169, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.15. URL https://aclanthology.org/2023.wmt-1.15.

Xu, H., Van Durme, B., and Murray, K. BERT, mBERT, or BiBERT? a study on contextualized embeddings for neural machine translation. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6663–6675, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.534. URL https://aclanthology.org/2021.emnlp-main.534.

Xu, H., Kim, Y. J., Sharaf, A., and Awadalla, H. H. A paradigm shift in machine translation: Boosting translation performance of large language models, 2023.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. mT5: A massively multilingual pre-trained text-to-text transformer. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y. (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41.

Yang, W., Li, C., Zhang, J., and Zong, C. Bigtrans: Augmenting large language models with multilingual translation capability over 100 languages. *arXiv preprint arXiv:2305.18098*, 2023.

Zeng, J., Meng, F., Yin, Y., and Zhou, J. Tim: Teaching large language models to translate with comparison. *arXiv preprint arXiv:2307.04408*, 2023.

Zhang, S., Fang, Q., Zhang, Z., Ma, Z., Zhou, Y., Huang, L., Bu, M., Gui, S., Chen, Y., Chen, X., et al. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. *arXiv preprint arXiv:2306.10968*, 2023.

Zhu, W., Liu, H., Dong, Q., Xu, J., Kong, L., Chen, J., Li, L., and Huang, S. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*, 2023a.

Zhu, W., Lv, Y., Dong, Q., Yuan, F., Xu, J., Huang, S., Kong, L., Chen, J., and Li, L. Extrapolating large language models to non-english by aligning languages. *arXiv preprint arXiv:2308.04948*, 2023b.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A. Comprehensive Results of WMT'21 And WMT'22

We show the comprehensive results for en→xx In Table 9 and xx→en in Table 10. In this section, our study additionally includes results from recently released LLM-based translators, including Bayling-13B (Zhang et al., 2023), BigTranslate (Yang et al., 2023), ALMA-13B-LoRA (Xu et al., 2023), the zero-shot performances of LLaMA-1-13B (Touvron et al., 2023a) and LLaMA-2-13B (Touvron et al., 2023b). We also compare these with the most advanced current translation models, such as WMT competition winners, GPT-4, GPT-3.5-text-davinci-003, Google Translate, and NLLB-3.3B. Importantly, we also present the performance of **ALMA-7B-R** here, which is fine-tuning on AMLA-7B-LoRA with CPO method. Except for reference-free evaluation, we also report two commonly used reference-based metrics, sacreBLEU (Post, 2018; Papineni et al., 2002) and COMET-22 (Unbabel/wmt22-comet-da) (Rei et al., 2022), in our analysis.

Introducing ALMA-7B-R In this study, we extend the ALMA-13B-R training methodology to a 7B model size, specifically fine-tuning ALMA-7B-LoRA using the CPO method with the same preference data as ALMA-13B-R. Consistent with our findings from ALMA-13B-R, the application of CPO significantly enhances performance. The ALMA-7B-R model surpasses ALMA-7B-LoRA by an average of 2.54 KIWI-XXL points and 1.44 XCOMET points for en→xx translations, and by 0.89 KIWI-XXL points and 3.69 XCOMET points for xx→en.

Comparing with Advanced Translation Models Our model, ALMA-13B-R, is benchmarked against the most advanced current models, demonstrating performance comparable to GPT-4 and WMT winners. It surpasses leading commercial translation tools such as Google Translate in many cases and top multilingual translation models like NLLB and GPT-3.5.
Stop Using BLEU BLEU, a metric extensively utilized for decades, often diverges from neural-based and reference-free metrics, a phenomenon also observed in previous studies (Xu et al., 2023; Freitag et al., 2023). For instance, WMT competition winners often exhibit superior performance according to BLEU (or COMET-22), yet this is not corroborated by reference-free models. A case in point is the WMT winners scoring an exceptionally high 64.14 BLEU in cs→en translations, significantly outperforming other models by 20 BLEU points. However, reference-free evaluations suggest these translations are inferior to those generated by our models and GPT-4. We hypothesize that this discrepancy may arise from WMT models being trained on domain-specific data closely related to the WMT test set, leading to high lexical matches but lacking semantic depth as evaluated by neural-based metrics. While BLEU scores are effective for assessing basic functionality in weaker models, their utility diminishes with advanced translation models capable of generating diverse translations. In such contexts, relying solely on BLEU for evaluation appears increasingly outdated.

Towards Reference-Free Metrics Neural-based, reference-dependent metrics like COMET-22 demonstrate greater consistency with reference-free metrics and robustness compared to BLEU. For instance, with COMET-22, our models show significant improvements like other reference-free models over ALMA-13B-LoRA and comparable performance to GPT-4, e.g., 87.74 (Ours) vs. 87.68 (GPT-4) when en→xx. However, it is important to note that, according to reference-free metrics, gold references are often inferior to system-generated translations, potentially indicating quality issues in the references that could impact COMET-22 evaluations. Consequently, inconsistencies still exist between COMET-22 and reference-free models like XCOMET. For example, XCOMET rates our ALMA+CPO model on average higher than WMT winners (89.11 vs. 87.13), while COMET-22 favors WMT winners (85.21 vs. 85.60). In line with the recommendations in Freitag et al. (2023), we advocate for the use of reference-free models to circumvent the potential quality issues of references.

Table 9: The full results in en→xx for WMT'21 and WMT'22 including both reference-free and reference-based metrics. **Bold** numbers denote the highest scores across all systems. Deep green boxes highlight improvements exceeding 1 point over the ALMA model after fine-tuning with preference data, while more modest gains under 1 point are shown in shallow green boxes. Decreases in performance are marked with red boxes. The asterisk (*) indicates that we directly utilized the reported translation outputs from Zhang et al. (2023) for evaluation purposes. Consequently, some baseline results for the is language are omitted in these instances.

	de					cs				
	BLEU	COMET-22	KIWI-22	KIWI-XXL	XCOMET	BLEU	COMET-22	KIWI-22	KIWI-XXL	XCOMET
Gold Reference	-	-	82.67	84.01	**97.85**	-	-	83.19	81.83	90.27
WMT Winners	**38.39**	87.21	83.56	83.70	96.99	**45.92**	**91.86**	**85.31**	**87.27**	**94.38**
GPT-4	34.58	87.29	83.48	**84.91**	97.56	33.74	90.81	84.81	85.35	93.48
GPT-3.5-text-davinci-003	31.88	85.61	82.75	81.71	96.35	31.31	88.57	82.93	78.84	89.91
Google Translate*	37.44	**88.01**	**84.03**	85.33	97.60	48.10	91.28	84.55	82.80	91.94
NLLB-3.3B*	34.04	86.24	83.38	82.47	96.25	36.34	89.90	84.20	81.77	91.57
LLaMA-1-13B	22.27	80.62	77.50	70.53	92.93	16.83	78.43	72.17	55.16	72.53
LLaMA-2-13B	13.69	75.55	68.33	55.98	90.81	0.87	68.57	61.38	42.67	74.26
Bayling-13B*	25.59	82.70	80.01	74.69	94.50	16.40	78.22	72.49	53.70	75.92
BigTranslate	22.13	78.62	75.40	67.45	90.22	20.57	80.11	73.53	60.27	73.73
ALMA-7B-LoRA	30.16	85.45	82.19	80.70	96.49	30.17	89.05	83.27	82.06	90.82
+ SFT on preferred data	29.00	85.42	82.3	80.44	96.26	30.64	89.11	83.46	81.28	90.26
+ DPO	28.87	85.19	82.02	80.02	96.22	28.87	88.78	82.96	81.03	90.12
+ CPO (Ours, ALMA-7B-R)	26.23	86.06	82.97	82.77	97.11	25.19	89.61	84.32	84.81	91.91
ALMA-13B-LoRA	31.47	85.62	82.62	81.64	96.49	32.38	89.79	84.14	84.24	92.38
+ SFT on preferred data	30.39	86.01	82.75	81.85	96.67	31.60	89.91	84.14	83.46	91.99
+ DPO	30.50	85.31	82.40	81.20	96.40	30.88	89.53	83.86	83.45	91.68
+ CPO (Ours, ALMA-13B-R)	27.72	86.40	83.28	84.25	97.48	26.32	90.29	84.99	87.06	93.61

	is					zh				
	BLEU	COMET-22	KIWI-22	KIWI-XXL	XCOMET	BLEU	COMET-22	KIWI-22	KIWI-XXL	XCOMET
Gold Reference	-	-	80.51	85.20	91.52	-	-	80.92	81.70	90.42
WMT Winners	**33.28**	86.75	81.77	84.94	91.61	**44.87**	86.69	82.04	81.13	91.14
GPT-4	24.68	85.08	81.03	81.21	90.00	44.41	86.51	81.73	81.53	90.79
GPT-3.5-text-davinci-003	15.89	76.28	73.96	58.72	72.67	38.36	85.76	81.26	80.4	90.07
Google Translate*	-	-	-	-	-	49.96	87.37	82.23	80.62	90.27
NLLB-3.3B*	-	-	-	-	-	32.52	81.57	75.73	67.14	82.04
LLaMA-1-13B	1.43	36.78	36.59	3.44	23.89	16.85	70.91	64.82	47.92	67.73
LLaMA-2-13B	2.36	38.47	31.07	4.60	36.21	30.00	79.70	74.09	65.06	81.06
Bayling-13B*	-	-	-	-	-	37.90	84.63	79.94	76.34	87.44
BigTranslate	2.08	37.40	39.29	9.39	26.77	19.17	74.11	65.96	55.44	72.69
ALMA-7B-LoRA	25.19	85.44	81.12	81.51	89.94	36.47	84.87	79.50	77.14	88.11
+ SFT on preferred data	24.26	85.19	80.87	80.25	89.15	37.12	85.36	80.38	78.16	88.34
+ DPO	24.52	85.20	80.79	80.42	88.97	35.22	84.73	79.42	76.96	87.72
+ CPO (Ours, ALMA-7B-R)	21.13	85.80	80.93	82.35	89.63	31.19	85.89	81.42	81.79	89.55
ALMA-13B-LoRA	26.68	86.08	81.71	83.31	91.20	39.84	85.96	80.82	79.96	89.92
+ SFT on preferred data	25.26	85.77	81.48	82.11	90.30	39.10	85.99	81.25	80.51	90.18
+ DPO	25.87	85.86	81.43	82.66	90.33	38.85	85.85	80.74	79.64	89.58
+ CPO (Ours, ALMA-13B-R)	22.88	**86.85**	**82.18**	**85.68**	**91.93**	34.06	**86.86**	**82.25**	**84.32**	**92.03**

	ru					Avg.				
	BLEU	COMET-22	KIWI-22	KIWI-XXL	XCOMET	BLEU	COMET-22	KIWI-22	KIWI-XXL	XCOMET
Gold Reference	-	-	82.96	84.62	94.17	-	-	82.05	83.47	92.85
WMT Winners	32.44	89.51	**84.35**	87.01	94.79	**38.98**	**88.40**	**83.41**	84.81	93.78
GPT-4	28.74	88.71	83.64	86.15	94.3	33.23	87.68	82.94	83.83	93.23
GPT-3.5-text-davinci-003	27.53	86.64	82.283	80.28	91.49	28.99	84.57	80.64	75.99	88.10
Google Translate*	**35.02**	**88.91**	83.75	84.26	93.50	-	-	-	-	-
NLLB-3.3B*	30.13	87.51	83.35	82.31	92.07	-	-	-	-	-
LLaMA-1-13B	18.46	79.16	74.26	64.72	86.32	15.17	69.18	65.07	48.35	68.68
LLaMA-2-13B	0.59	63.84	56.78	38.53	84.94	9.50	65.23	58.33	41.37	73.46
Bayling-13B*	12.76	71.01	67.89	54.62	85.63	-	-	-	-	-
BigTranslate	16.14	75.13	69.22	54.27	76.92	16.02	69.07	64.68	49.36	68.07
ALMA-7B-LoRA	26.93	87.05	82.72	82.60	92.98	29.78	86.37	81.76	80.80	91.67
+ SFT on preferred data	26.23	86.88	82.47	81.79	92.57	29.45	86.39	81.90	80.38	91.32
+ DPO	25.94	86.70	82.20	81.61	92.53	28.68	86.12	81.48	80.01	91.11
+ CPO (Ours, ALMA-7B-R)	23.31	87.86	83.45	84.97	94.15	25.41	87.04	82.62	83.34	92.47
ALMA-13B-LoRA	28.96	87.53	83.10	84.17	93.79	31.87	87.00	82.48	82.66	92.76
+ SFT on preferred data	28.15	87.66	83.23	84.15	93.54	30.90	87.07	82.57	82.42	92.54
+ DPO	28.27	87.30	82.94	83.40	93.25	30.87	86.77	82.27	82.07	92.25
+ CPO (Ours, ALMA-13B-R)	24.15	88.30	83.98	**87.37**	**95.22**	27.03	87.74	83.34	**85.74**	**94.05**

Table 10: The full results in xx→en for WMT'21 and WMT'22 including both reference-free and reference-based metrics. The usage of color and boldface are the same in Table 9. The asterisk (*) indicates that we directly utilized the reported translation outputs from Zhang et al. (2023) for evaluation purposes. Consequently, some baseline results for the is language are omitted in these instances.

	de					cs				
	BLEU	COMET-22	KIWI-22	KIWI-XXL	XCOMET	BLEU	COMET-22	KIWI-22	KIWI-XXL	XCOMET
Gold Reference	-	-	78.74	78.56	88.82	-	-	82.08	83.11	84.60
WMT Winners	33.34	85.04	81.38	83.59	93.74	**64.14**	**89.00**	82.47	82.53	85.65
GPT-4	32.41	85.35	**81.50**	**84.58**	**94.47**	46.86	87.26	82.52	83.55	**88.48**
GPT-3.5-text-davinci-003	30.78	84.79	81.24	83.97	92.78	44.51	86.16	82.02	82.19	83.51
Google Translate*	**33.25**	84.78	81.36	83.74	93.71	49.40	86.95	82.60	81.99	86.74
NLLB-3.3B*	29.46	83.43	80.98	82.04	91.26	49.05	85.92	81.72	80.27	82.94
LLaMA-1-13B	29.66	82.42	78.77	77.98	89.99	36.05	81.57	77.72	70.80	73.71
LLaMA-2-13B	31.06	83.01	79.47	79.27	91.10	40.02	83.27	79.29	74.21	78.50
Bayling-13B*	27.26	83.03	79.88	80.02	89.84	33.81	81.65	78.04	71.44	71.68
BigTranslate	25.16	81.54	78.24	77.73	86.79	34.81	82.02	77.91	72.69	71.38
ALMA-7B-LoRA	29.56	83.95	80.63	82.58	92.35	43.49	85.93	81.32	81.42	81.34
+ SFT on preferred data	30.51	84.39	80.86	82.72	93.19	44.44	86.17	81.97	81.95	84.55
+ DPO	29.38	84.02	80.63	82.47	92.26	42.60	85.87	81.33	81.30	81.10
+ CPO (Ours, ALMA-7B-R)	30.52	84.61	81.13	83.11	93.85	42.92	86.29	82.16	82.29	85.76
ALMA-13B-LoRA	31.14	84.56	81.14	83.57	93.30	45.28	86.47	81.96	82.97	83.95
+ SFT on preferred data	31.80	84.83	81.36	83.98	93.84	46.17	86.83	82.36	83.15	86.67
+ DPO	30.99	84.51	81.13	83.52	93.25	44.95	86.36	81.82	82.69	83.84
+ CPO (Ours, ALMA-13B-R)	30.89	**84.95**	**81.50**	83.97	94.20	44.39	86.85	**82.63**	**83.75**	88.03

	is					zh				
	BLEU	COMET-22	KIWI-22	KIWI-XXL	XCOMET	BLEU	COMET-22	KIWI-22	KIWI-XXL	XCOMET
Gold Reference	-	-	80.88	85.04	76.16	-	-	77.09	74.19	90.70
WMT Winners	**41.6**	86.98	81.39	85.60	78.14	**33.49**	81.02	77.66	73.28	87.2
GPT-4	41.29	87.21	81.49	**85.90**	**81.11**	23.82	**82.46**	**79.33**	**77.65**	**92.06**
GPT-3.5-text-davinci-003	31.88	82.13	78.72	77.53	66.44	24.98	81.62	78.91	76.64	90.92
Google Translate*	-	-	-	-	-	28.60	80.82	77.87	74.27	87.69
NLLB-3.3B*	-	-	-	-	-	21.08	76.93	75.40	68.83	84.43
LLaMA-1-13B	11.01	60.82	57.76	30.38	20.87	16.81	74.32	70.93	62.37	80.13
LLaMA-2-13B	15.77	66.35	63.91	42.75	28.03	21.81	78.10	75.09	70.31	85.68
Bayling-13B*	-	-	-	-	-	20.10	77.72	75.08	68.32	86.51
BigTranslate	6.45	54.65	50.55	18.77	17.44	14.94	75.11	71.94	65.25	85.00
ALMA-7B-LoRA	35.64	86.09	80.57	84.65	75.02	23.64	79.78	76.81	73.65	83.94
+ SFT on preferred data	38.58	86.47	81.09	85.23	78.87	23.19	80.50	77.74	74.91	89.81
+ DPO	35.25	85.96	80.53	84.44	75.19	23.20	79.91	76.83	73.51	89.22
+ CPO (Ours, ALMA-7B-R)	38.64	86.66	81.24	85.13	79.14	22.45	80.95	78.47	75.72	90.74
ALMA-13B-LoRA	36.95	86.42	80.90	85.49	76.68	25.46	80.21	77.32	74.41	89.88
+ SFT on preferred data	39.60	86.88	81.32	85.61	80.20	24.54	81.08	78.32	76.03	90.65
+ DPO	36.16	86.30	80.89	85.22	76.09	25.17	80.42	77.50	74.50	89.94
+ CPO (Ours, ALMA-13B-R)	39.67	**87.14**	**81.57**	85.73	80.49	23.23	81.64	79.24	77.17	91.65

	ru					Avg.				
	BLEU	COMET-22	KIWI-22	KIWI-XXL	XCOMET	BLEU	COMET-22	KIWI-22	KIWI-XXL	XCOMET
Gold Reference	-	-	80.74	79.59	88.56	-	-	79.91	80.10	85.77
WMT Winners	**45.18**	**85.95**	81.71	80.97	90.91	**43.55**	**85.60**	80.92	81.19	87.13
GPT-4	41.09	85.87	81.57	81.34	90.95	37.09	85.63	81.28	**82.60**	**89.41**
GPT-3.5-text-davinci-003	38.52	84.8	81.14	79.95	89.29	34.13	83.90	80.41	80.06	84.59
Google Translate*	43.66	84.81	81.02	79.66	89.40	-	-	-	-	-
NLLB-3.3B*	40.12	83.95	80.87	78.37	87.85	-	-	-	-	-
LLaMA-1-13B	34.65	81.90	78.29	74.37	84.13	25.64	76.21	72.69	63.18	69.77
LLaMA-2-13B	36.50	82.91	79.14	76.50	86.12	29.03	78.73	75.38	68.61	73.89
Bayling-13B*	33.94	82.07	78.72	74.45	83.28	-	-	-	-	-
BigTranslate	29.06	78.10	74.35	66.46	72.78	22.08	74.28	70.60	60.18	66.68
ALMA-7B-LoRA	39.21	84.84	80.94	80.19	88.50	34.31	84.12	80.05	80.50	84.23
+ SFT on preferred data	39.06	85.00	81.05	80.47	89.54	35.16	84.51	80.54	81.06	87.19
+ DPO	38.40	84.71	80.83	80.04	88.34	33.77	84.09	80.03	80.35	85.22
+ CPO (Ours, ALMA-7B-R)	38.42	85.11	81.34	80.69	90.10	34.59	84.72	80.87	81.39	87.92
ALMA-13B-LoRA	40.27	85.27	81.31	81.05	89.89	35.82	84.59	80.53	81.50	86.74
+ SFT on preferred data	40.55	85.44	81.46	81.17	90.65	36.53	85.01	80.96	81.99	88.40
+ DPO	39.12	85.14	81.19	80.88	89.76	35.28	84.55	80.51	81.36	86.58
+ CPO (Ours, ALMA-13B-R)	39.06	85.45	**81.72**	**81.54**	**91.18**	35.45	85.21	**81.33**	82.43	89.11

B. Prompts for Translations

Adhering to the prompt format for translation as utilized by Hendy et al. (2023) for GPT models, we employ the same prompt for GPT-4 in our study. Similarly, we use the same prompt employed by Xu et al. (2023) for ALMA models. Prompts are depicted in Figure 4.

GPT-4 Prompt

System:
You are a helpful translator and only output the result.

User:
Translate this from <source language> to <target language>, <source language>:
<source sentence>
<target language>:

ALMA Prompt

Translate this from <source language> to <target language>:
<source language>: <source sentence>
<target language>:

Figure 4: The prompts employed for GPT-4 and ALMA models to perform translations.

C. Simplifying The Re-Written DPO Loss Function

In this section, we provide a detailed explanation of the approximation used to optimize the loss function from Equation 3 to Equation 6. We begin by expanding the Sigmoid function in Equation 3:

$$\begin{aligned}
\mathcal{L}(\pi_\theta;\pi_w) &= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\sigma\Big(\beta\log\pi_\theta(y_w|x)-\beta\log\pi_\theta(y_l|x)+\beta\log\pi_w(y_l|x)\Big)\Big]\\
&= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\Big(\frac{1}{1+e^{-\beta\log\pi_\theta(y_w|x)+\beta\log\pi_\theta(y_l|x)-\beta\log\pi_w(y_l|x)}}\Big)\Big]\\
&= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\Big(\frac{1}{1+\frac{\pi_\theta(y_l|x)^\beta}{\pi_\theta(y_w|x)^\beta\cdot\pi_w(y_l|x)^\beta}}\Big)\Big]\\
&= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\pi_\theta(y_w|x)^\beta+\log\pi_w(y_l|x)^\beta-\log\Big(\pi_\theta(y_w|x)^\beta\cdot\pi_w(y_l|x)^\beta+\pi_\theta(y_l|x)^\beta\Big)\Big]. \qquad (9)
\end{aligned}$$

Given that π_w is a fixed model and $\log\pi_w(y_l|x)^\beta$ does not participate in gradient calculations or parameter updates, the above loss function can be simplified by omitting the term $\log\pi_w(y_l|x)^\beta$:

$$\mathcal{L}(\pi_\theta;\pi_w) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\pi_\theta(y_w|x)^\beta-\log\Big(\pi_\theta(y_w|x)^\beta\cdot\pi_w(y_l|x)^\beta+\pi_\theta(y_l|x)^\beta\Big)\Big]. \qquad (10)$$

Considering that $0\le\pi_w(y_l|x)\le 1$, the loss can be upper bounded as follows:

$$\begin{aligned}
\mathcal{L}(\pi_\theta;\pi_w) &\le -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\pi_\theta(y_w|x)^\beta-\log\Big(\pi_\theta(y_w|x)^\beta\cdot 1+\pi_\theta(y_l|x)^\beta\Big)\Big]\\
&= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\sigma\Big(\beta\log\pi_\theta(y_w|x)-\beta\log\pi_\theta(y_l|x)\Big)\Big]. \qquad (11)
\end{aligned}$$

Therefore, minimizing the loss function as presented in Equation 3 can be effectively approximated by minimizing its upper bound, as illustrated in Equation 6 and 11.

D. Details And Influence of Human-Labeled Preference Data

TL;DR: Our analysis indicates that our human-labeled data has a relatively minimal impact, probably due to a high proportion of tied translations and potential human bias in the evaluation process.

D.1. Data Construction Details

The human-labeled dataset we used is pair-wise and differs from the triplet format of our main dataset. It focuses exclusively on two language directions, en→de and en→zh, resulting in an additional 2K sentences. The English source sentences, selected from Wikipedia, undergo a filtering process to remove time stamps and URLs. Each sentence is translated using Google Translate and GPT-4, with human evaluators then assigning their preference between these two translations. The distribution of preferences, indicating the number of times translations from Google or GPT-4 were favored or instances where they tied, is detailed in Table 11.

Table 11: The statistic of how many translations win or tie by each system evaluated by human.

	Google Wins	GPT-4 Wins	Ties
en→de	418	435	203
en→zh	362	412	282

D.2. Influence on Performance

Given that our model operates in a many-to-many translation format and the additional data is specific to only de and zh directions, we anticipate changes in performance when translating into these languages, but not in others. To assess the impact of the human-labeled data, we conducted a comparison between models exclusively fine-tuned on triplet data and those fine-tuned on both triplet and human-labeled data. The training approach remained consistent, utilizing the ALMA-13B-LoRA model fine-tuned via CPO. It's important to note that tied data were excluded from this analysis due to their lack of clear preference.

Results and Analysis We show the detailed results for en→xx and xx→en in Table 12 and 13, respectively. The inclusion of human-labeled preference data does not significantly enhance overall translation performance. For en→zh, marginal improvements are observed, though they are minimal. Conversely, for en→de, a slight decline in performance is noted. In summary, the addition of human-labeled data shows no substantial difference in the en→xx direction, and a minor decrease in performance for xx→en on average. We hypothesize that the limited impact of these human-labeled data may stem from a high proportion of tied evaluations and potential human bias in the evaluation process. For instance, there are instances where the author consider GPT-4's translations to be superior, while human evaluators favor those produced by Google Translate.

Table 12: A comparison of translation performance when utilizing solely triplet data versus a combination of triplet data and human-labeled data (our original setup) in the en→xx direction. The **bold** number indicates superior performance. There is not obvious performance difference adding our human-labeled data.

Dataset	de			cs			is		
	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET
Only Triplet Data	**83.43**	**84.63**	**97.56**	84.97	**87.24**	93.50	82.05	85.37	91.83
Triplet Data + Human-Labeled Data	83.28	84.25	97.48	**84.99**	87.06	**93.61**	**82.18**	**85.68**	**91.93**
Dataset	zh			ru			Avg.		
	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET
Only Triplet Data	82.15	84.08	91.59	**84.05**	**87.43**	**95.26**	83.33	**85.75**	93.95
Triplet Data + Human-Labeled Data	**82.25**	**84.32**	**92.03**	83.98	87.37	95.22	**83.34**	85.74	**94.05**

Table 13: A comparison of translation performance when utilizing solely triplet data versus a combination of triplet data and human-labeled data (our original setup) in the en→xx direction. The **bold** number indicates superior performance. Interestingly, the inclusion of our human-labeled data results in a slight decrease in average performance.

Dataset	de			cs			is		
	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET
Only Triplet Data	**81.57**	**84.25**	**94.32**	**82.68**	83.70	87.97	**81.63**	**85.87**	80.89
Triplet Data + Human-Labeled Data	81.50	83.97	94.20	82.63	**83.75**	**88.03**	81.57	85.73	**80.49**
Dataset	zh			ru			Avg.		
	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET
Only Triplet Data	**79.34**	**77.31**	**91.76**	**81.76**	**81.63**	**91.34**	**81.40**	**82.55**	**89.26**
Triplet Data + Human-Labeled Data	79.24	77.17	91.65	81.72	81.54	91.18	81.33	82.43	89.11

E. WMT Winner Systems

E.1. Systems For WMT'21 And WMT'22

The WMT competition winners for each direction as reported in WMT'21 and WMT'22 correspond to those used by Hendy et al. (2023). For more detailed information, we direct readers to this paper.

E.2. Systems For WMT'23

For the de↔en and zh↔en language pairs, we selected the translation systems that attained the highest human rankings based on source-based Direct Assessment and Scalar Quality Metrics (DA+SQM). For de↔ru, in the absence of human rankings for these directions in Kocmi et al. (2023), we opted for the model with the highest COMET-22 scores as reported in Kocmi et al. (2023). Details about these models are available in Table 14.

Table 14: The list of WMT'23 winners served for each language direction.

Systems	Language Pair
ONLINE-B	en-de
ONLINE-A	de-en
Lan-BridgeMT (Wu & Hu, 2023)	en-zh
Lan-BridgeMT (Wu & Hu, 2023)	zh-en
ONLINE-G	en-ru
ONLINE-Y	ru-en

F. Full Results of WMT'23

The comprehensive results of WMT'23 are presented in Table 15. Similar to its performance in WMT'21 and WMT'22, ALMA-13B-R performs best on average among the SoTA translation models.

Table 15: The full results of WMT'23. The highest score among all systems are bold.

	de→en			zh→en			ru→en		
	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET
Gold Reference	78.93	75.96	84.23	74.46	68.80	83.51	79.46	77.84	83.60
WMT Winners	79.37	76.18	84.35	**80.17**	**79.53**	92.25	80.88	79.21	86.22
TowerInstruct	79.67	77.60	86.28	79.84	78.13	91.75	80.85	80.03	87.76
ALMA-13B-LoRA	79.36	76.79	85.07	78.83	76.71	90.73	80.79	80.14	86.94
+ CPO (Ours, ALMA-13B-R)	**79.87**	**77.69**	**86.62**	80.01	78.42	**92.36**	**81.11**	**80.95**	**88.75**
	en→de			en→zh			en→ru		
	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET	KIWI-22	KIWI-XXL	XCOMET
Gold Reference	80.12	**77.93**	88.91	79.60	73.47	86.15	79.87	79.36	91.41
WMT Winners	**80.80**	77.26	87.94	79.70	74.20	87.24	**82.51**	79.95	91.41
TowerInstruct	80.13	75.34	86.55	80.03	74.85	86.74	81.33	77.14	89.59
ALMA-13B-LoRA	78.79	73.40	85.61	78.92	72.95	85.13	80.21	76.02	89.48
+ CPO (Ours, ALMA-13B-R)	79.85	77.05	**89.79**	**80.48**	**78.17**	**88.34**	81.97	**81.52**	**92.56**