

CLewR: Curriculum Learning with Restarts for Machine Translation Preference Learning

Alexandra Dragomir¹, Florin Brad¹, Radu Tudor Ionescu^{2,◇}

¹Bitdefender, Bucharest, Romania

²Department of Computer Science, University of Bucharest, Bucharest, Romania

◇raducu.ionescu@gmail.com

Abstract

Large language models (LLMs) have demonstrated competitive performance in zero-shot multilingual machine translation (MT). Some follow-up works further improved MT performance via preference optimization, but they leave a key aspect largely underexplored: the order in which data samples are given during training. We address this topic by integrating curriculum learning into various state-of-the-art preference optimization algorithms to boost MT performance. We introduce a novel curriculum learning strategy with restarts (CLewR), which reiterates easy-to-hard curriculum multiple times during training to effectively mitigate the catastrophic forgetting of easy examples. We demonstrate consistent gains across several model families (Gemma2, Qwen2.5, Llama3.1) and preference optimization techniques. We publicly release our code at <https://github.com/alexandra-dragomir/CLewR>.

1 Introduction

Large language models (LLMs) have enabled zero-shot approaches in multilingual machine translation (MT) (Touvron et al., 2023). Methods for improving the MT abilities of LLMs can be broadly categorized into pre-training and post-training approaches. The former typically employ continual pre-training over large-scale monolingual or high-quality parallel data (Alves et al., 2024; Cui et al., 2025; Xu et al., 2024a). In contrast, post-training approaches aim to improve translation quality by employing preference optimization techniques, such as Direct Preference Optimization (DPO) (Rafailov et al., 2023), to distinguish high-quality translations from low-quality ones. Building on this line of work, Xu et al. (2024b) proposed Contrastive Preference Optimization (CPO), a reference-free technique that assesses pair distances based on log-probability differences only. More recently, Xu et al. (2025) introduced Adaptive Rejection Preference Optimization

(ARPO), which further improves CPO by incorporating an adaptive penalty for the unpreferred term.

Despite the significant advances in preference optimization (PO) techniques (Rafailov et al., 2023; Xu et al., 2025, 2024b), a key factor that can significantly influence performance remains underexplored: the order in which data samples are processed during training. This aspect is central to *curriculum learning* (Bengio et al., 2009), a paradigm that studies how models can learn from easy to hard. The survey of Soviany et al. (2022) explains that an easy-to-hard learning can be created by manipulating distinct factors, namely the data (Chang et al., 2021; Jarca et al., 2024; Nagatsuka et al., 2023), the model (Croitoru et al., 2025b; Sinha et al., 2020) or the target task (Liu et al., 2020a; Narvekar et al., 2016). Organizing the samples in a certain order falls under the umbrella of *data-level curriculum*. In the realm of data-level curriculum, researchers have explored both easy-to-hard and hard-to-easy data organizations, the latter being known as anti-curriculum (Ankner et al., 2024; Florensa et al., 2017; Jarca et al., 2025). Regardless of the data organization, several recent studies showed that curriculum learning can play an important role in various tasks, e.g. natural language inference (Poesina et al., 2024), intent detection (Gong et al., 2021), question answering (Liu et al., 2018), image classification (Liu et al., 2022), model pre-training (Madan et al., 2024; Nagatsuka et al., 2023), etc. Curriculum learning has also been applied in neural machine translation (NMT) (Kocmi and Bojar, 2017; Liu et al., 2020b; Platanios et al., 2019; Zhan et al., 2021), but contributions in this area predate the era of LLMs, making them hard or impossible to adapt to the novel “pre-training then fine-tuning” paradigm. With the emergence of preference optimization techniques applied during the fine-tuning stage (Rafailov et al., 2023; Xu et al., 2025, 2024b), some recent works (Croitoru et al., 2025a; Pattnaik

et al., 2024) have integrated curriculum learning into DPO. However, such techniques do not explicitly address catastrophic forgetting (Kirkpatrick et al., 2017), a problem that occurs when samples learned at the beginning are forgotten by the model by the end of the training process, eventually degrading performance.

To this end, we propose a novel data-level curriculum learning framework for MT, in which the easy-to-hard training is restarted at every epoch. Our curriculum learning strategy with restarts (CLewR) is natively designed to mitigate catastrophic forgetting by iterating through all samples in every training epoch. We empirically demonstrate that CLewR leads to consistent performance gains in MT across several state-of-the-art preference optimization methods (DPO, CPO, ARPO) and LLM families (Gemma2, Qwen2.5, Llama3.1). Our results show that CLewR not only enhances highly competitive preference optimization methods, but also surpasses another competitor based on curriculum learning, namely CurriDPO (Pattanaik et al., 2024).

In summary, our contribution is threefold:

- We propose curriculum learning with restarts (CLewR), a novel method for preference optimization in MT, where the easy-to-hard curriculum is restarted at every epoch to avoid catastrophic forgetting.
- While previous work enhanced DPO with curriculum (Pattanaik et al., 2024), we introduce curriculum learning to newer preference optimization algorithms, namely CPO and ARPO.
- We show that our method outperforms competing curriculum approaches and consistently improves performance across multiple model families (Gemma2, Llama3.1, Qwen2.5) and preference optimization algorithms (DPO, CPO, ARPO).

2 Method

CLewR. We propose a data-level curriculum strategy, named CLewR, which is tailored to preference optimization in MT. We formally present our curriculum strategy in Algorithm 1. Training preference triplets of the form (x, y_w, y_l) are ordered (in step 9) based on a similarity score $s(y_w, y_l)$ between the chosen (winning) y_w and rejected (losing) y_l translations. More precisely, the easiness of a pair of translations is defined as the similarity difference between the preferred and rejected

Algorithm 1 CLewR Preference Optimization

```

1: Input: initial policy  $\pi_\theta$ , training triplets
    $\{(x_i, y_i^w, y_i^l)\}_{i=1}^N$ , learning rate  $\mu$ .
2: for  $i = 1$  to  $N$  do
3:    $b \leftarrow \text{BLEU}(y_i^w, y_i^l)$ 
4:    $c \leftarrow \text{COMET}(y_i^w, y_i^l)$ 
5:    $m \leftarrow \text{METEOR}(y_i^w, y_i^l)$ 
6:    $\hat{b}, \hat{c}, \hat{m} \leftarrow \text{normalize}_{(0,1)}(b, c, m)$ 
7:    $s_i \leftarrow \frac{1}{3}(\hat{b} + \hat{c} + \hat{m})$ 
8: end for
9:  $\mathcal{I} \leftarrow \text{argsort}_\uparrow(\{s_i\}_{i=1}^N)$ 
10: for epoch = 1 to  $E$  do
11:   for all batches  $\mathcal{B} \subset \mathcal{I}$  do
12:      $\mathcal{L}_{\text{PO}} \leftarrow \text{loss}(x_{\mathcal{B}}, y_{\mathcal{B}}^w, y_{\mathcal{B}}^l, \pi_\theta)$ 
13:      $\theta \leftarrow \text{optimize}(\theta, \mu, \nabla_\theta \mathcal{L}_{\text{PO}})$ 
14:   end for
15: end for
16: Output: optimized model  $\pi_\theta$ 

```

translation, i.e. a high difference corresponds to an easy pair and a low difference to a hard pair. The similarity score is given by the average (step 7) of multiple MT metrics, thus making CLewR suitable for translation: BLEU (Papineni et al., 2002) (step 3), COMET-22 (Rei et al., 2022) (step 4), and METEOR (Banerjee and Lavie, 2005) (step 5) scores.

We emphasize that our method implicitly works with multiple correct reference translations for a given source sentence. By default, preference optimization techniques work with triplets of the form (x, y_w, y_l) . If the dataset includes k preferred outputs for the same input, we can build k preference optimization triplets. Then, CLewR can simply apply PO starting from the easier tuples (references that are most dissimilar to the rejected sample) to the more difficult tuples.

We provide examples of easy and hard preference samples in Appendix A.5. After sorting the triplets, training proceeds over a number of epochs (steps 10-15). At every epoch, the samples are divided into mini-batches (step 11) in the exact order established at step 9, i.e. there is no random shuffling involved. The easy-to-hard data permutation is reused at every epoch, which helps mitigate catastrophic forgetting. Learning is performed via a given PO method (steps 12-13). Note that fixing the order of samples in each epoch does not imply overfitting the order, i.e. the empirical risk does not depend on sample order. On the contrary, curriculum learning theory (Bengio et al., 2009) suggests

that organizing the samples in a meaningful order can improve training dynamics, potentially leading to faster convergence and/or better optima.

CLewR-z. For ARPO (Xu et al., 2025), we develop an alternative CLewR implementation called CLewR-z, where the curriculum score s (used in step 7 of Algorithm 1) is derived from the ARPO distance z . The ARPO objective modifies the CPO objective to incorporate an adaptive penalty term τ_θ , which controls the importance of the rejected term y_l :

$$\begin{aligned} \mathcal{L}_{\text{ARPO}} = & - \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \pi_\theta(y_w | x) \right. \right. \\ & \left. \left. - \tau_\theta(y_w, y_l) \beta \log \pi_\theta(y_l | x) \right) \right. \\ & \left. + \log \pi_\theta(y_w | x) \right]. \end{aligned} \quad (1)$$

The value of τ_θ measures the similarity between y_w and y_l , ranging from 0 to 1:

$$\tau_\theta(y_w, y_l) = \min(e^{\eta \cdot z_\theta(y_w, y_l)} - 1, 1), \quad (2)$$

where η is a hyperparameter that controls the impact of z_θ , and $z_\theta(y_w, y_l)$ encodes the distance between the chosen and rejected responses by measuring the absolute difference in log-likelihoods:

$$z_\theta(y_w, y_l) = \left| \frac{\log(\pi_\theta(y_w | x))}{|y_w|} - \frac{\log(\pi_\theta(y_l | x))}{|y_l|} \right|. \quad (3)$$

For curriculum learning, we employ $s = -z_\theta$ in step 7 of Algorithm 1. This version is called CLewR-z.

Enhanced ARPO. We further introduce an enhanced variant of ARPO by using a different distance function $z'_\theta(y_w, y_l)$ that also accounts for distances in the evaluation metric spaces. Specifically, we use:

$$z' = \eta_1 \cdot z_\theta + \eta_2 \cdot z_{\text{BLEU}} + \eta_3 \cdot z_{\text{COMET}}, \quad (4)$$

where z_θ is the original distance. For both metrics, z_{metric} is given by $1 - \frac{\text{metric}}{100}$ to normalize each metric to a $(0, 1)$ interval and have the same monotony as the original z_θ . Two dissimilar predictions result in low BLEU and COMET values, so z_{BLEU} and z_{COMET} will be high. Each z is multiplied by a corresponding scalar η , scaling them to similar intervals. We create multiple versions of enhanced ARPO by modifying the scalars η_1 , η_2 and η_3 . All ARPO versions are listed in Table 6.

3 Experimental Setup

Dataset. We test on the Flores-200 (Costa-Jussà et al., 2022) dataset. For generic LLMs, we use a group of six Romance languages. For MT-adapted models (GemmaX2), we use a group of three Ro-

mance languages, following Cui et al. (2025). We select Chinese to showcase generalization beyond Romance languages.

LLM backbones. We consider several candidate LLMs for preference tuning: LLama3.1-8B (Grattafiori et al., 2024), Qwen2.5-7B (Qwen et al., 2025), Gemma2-9B (Team et al., 2024), and GemmaX2-9B (Cui et al., 2025). We also consider X-ALMA (Xu et al., 2025) as a reference baseline, which is based on Llama2 (Touvron et al., 2023).

Preference optimization baselines. We train the models using three different PO algorithms: Direct Preference Optimization Positive (DPOP) (Pal et al., 2024) (an enhanced version of DPO (Rafailov et al., 2023)), CPO with behavior cloning (Xu et al., 2024b), and ARPO (Xu et al., 2025). We employ these algorithms with and without our CLewR. For ARPO, we integrate both CLewR and CLewR-z.

Another comparison is between CurriDPO (Pattnaik et al., 2024) and CLewR. For CurriDPO, we use the iterative approach based on the same SFT reference model.

Hyperparameter tuning. We fine-tune LoRA (Hu et al., 2022) adapters on top of the base models, using a rank of 64 for the LoRA matrices, a learning rate of $5 \cdot 10^{-5}$, a warmup ratio of 0.1, and a mini-batch size of 4. All models are optimized with Adam (Kingma and Ba, 2015). Gemma2, Qwen2.5 and Llama3.1 are each trained for three epochs. Since GemmaX2 is already tuned for MT, it is trained in one epoch. We use default values for other hyperparameters.

Training and evaluation protocol. We focus on translation to and from English for languages from the Romance language group used in X-ALMA (Xu et al., 2025): Catalan (ca), Galician (gl), Italian (it), Portuguese (pt), Romanian (ro) and Spanish (es). We use the X-ALMA¹ preference data (81K train and 7K validation) and select the triplets from the en→xx and xx→en language pairs, where xx belongs to the Romance language group. For Gemma2-9B, Qwen2.5-7B and LLama3.1-8B, we use all six languages, while for GemmaX2, we select {it, pt, es}, since the model was not adapted for MT on the other three languages.

Furthermore, to assess the generalization of our method beyond Romance languages, we also report results on a linguistically distant language pair. Specifically, we evaluate Gemma2-9B on Chinese (zh) using the X-ALMA preference dataset, con-

¹XALMA preference dataset.

Model	Test on 6 languages (Romance group)			
	en→xx		xx→en	
	BLEU ↑	COMET ↑	BLEU ↑	COMET ↑
X-ALMA (ARPO)	39.83	89.38	42.88	89.28
Gemma2-9B	19.85	61.92	34.98	85.57
+ DPOP	23.26 \pm 4.23	85.94 \pm 0.59	33.17 \pm 1.47	87.77 \pm 0.02
+ CurriDPO	21.81 \pm 0.26	85.64 \pm 0.34	28.33 \pm 2.71	87.27 \pm 0.20
+ CLewR	22.35 \pm 1.85	86.11 \pm 0.18	31.86 \pm 1.40	87.65 \pm 0.17
+ CPO	33.53 \pm 0.09	87.43 \pm 0.05	36.44 \pm 0.06	87.76 \pm 0.03
+ CLewR†	36.24 \pm 0.03	87.85 \pm 0.01	40.33 \pm 0.23	88.49 \pm 0.04
+ ARPO	35.37 \pm 0.02	88.12 \pm 0.04	37.68 \pm 0.63	88.30 \pm 0.10
+ CLewR†	36.63 \pm 0.05	88.12 \pm 0.01	40.44 \pm 0.06	88.56 \pm 0.02
+ CLewR-z†	36.50 \pm 0.02	88.12 \pm 0.03	39.87 \pm 0.34	88.62 \pm 0.06
+ ARPO-z′-V1	36.35 \pm 0.01	88.40 \pm 0.02	40.32 \pm 0.09	88.62 \pm 0.03
+ CLewR†	37.45 \pm 0.02	88.40 \pm 0.04	41.28 \pm 0.11	88.78 \pm 0.03
+ CLewR-z†	36.98 \pm 0.03	88.38 \pm 0.04	41.22 \pm 0.08	88.84 \pm 0.02
+ ARPO-z′-V2	36.46 \pm 0.05	88.43 \pm 0.03	40.10 \pm 0.16	88.58 \pm 0.05
+ CLewR†	37.45 \pm 0.04	88.37 \pm 0.05	41.09 \pm 0.21	88.75 \pm 0.05
+ CLewR-z†	37.08 \pm 0.02	88.38 \pm 0.01	41.45 \pm 0.08	88.87 \pm 0.02
Qwen2.5-7B	4.10	45.94	9.59	62.83
+ DPOP	24.43 \pm 0.88	82.90 \pm 0.24	33.06 \pm 0.76	87.42 \pm 0.12
+ CurriDPO	23.80 \pm 0.26	81.90 \pm 0.30	29.12 \pm 2.58	86.77 \pm 0.39
+ CLewR	23.59 \pm 1.63	82.74 \pm 0.41	30.63 \pm 1.97	87.10 \pm 0.27
+ CPO	27.68 \pm 0.03	83.93 \pm 0.06	31.66 \pm 0.03	86.36 \pm 0.03
+ CLewR†	30.05 \pm 0.02	84.15 \pm 0.06	36.28 \pm 0.10	87.33 \pm 0.05
+ ARPO	30.41 \pm 0.05	85.64 \pm 0.04	32.12 \pm 0.10	86.71 \pm 0.04
+ CLewR†	31.56 \pm 0.10	85.38 \pm 0.07	36.69 \pm 0.27	87.62 \pm 0.10
+ CLewR-z†	31.14 \pm 0.06	85.47 \pm 0.04	34.88 \pm 0.08	87.45 \pm 0.02
+ ARPO-z′-V1	31.08 \pm 0.14	85.67 \pm 0.01	35.12 \pm 0.91	87.37 \pm 0.29
+ CLewR†	32.68 \pm 0.02	85.81 \pm 0.01	37.18 \pm 0.08	87.60 \pm 0.03
+ CLewR-z†	32.68 \pm 0.05	85.93 \pm 0.05	36.55 \pm 0.01	87.56 \pm 0.02
+ ARPO-z′-V2	30.95 \pm 0.09	85.67 \pm 0.03	36.40 \pm 0.10	87.75 \pm 0.02
+ CLewR†	32.70 \pm 0.03	85.78 \pm 0.01	37.16 \pm 0.06	87.64 \pm 0.03
+ CLewR-z	32.67 \pm 0.04	85.93 \pm 0.02	36.15 \pm 0.08	87.51 \pm 0.02
LLaMA3.1-8B	1.70	46.00	1.84	46.30
+ DPOP	31.53 \pm 0.11	87.05 \pm 0.08	35.78 \pm 0.46	87.82 \pm 0.01
+ CurriDPO	32.46 \pm 0.36	86.94 \pm 0.06	36.01 \pm 0.50	87.75 \pm 0.05
+ CLewR	30.87 \pm 0.37	86.85 \pm 0.07	35.46 \pm 0.65	87.83 \pm 0.03
+ CPO	1.47 \pm 0.34	54.97 \pm 2.52	10.36 \pm 4.48	74.32 \pm 1.78
+ CLewR†	3.97 \pm 0.99	59.66 \pm 3.21	13.08 \pm 1.93	76.76 \pm 3.90
+ ARPO	33.50 \pm 0.20	87.48 \pm 0.05	35.92 \pm 0.56	87.75 \pm 0.11
+ CLewR†	34.49 \pm 0.05	87.44 \pm 0.06	37.99 \pm 0.07	87.94 \pm 0.04
+ CLewR-z†	34.33 \pm 0.09	87.57 \pm 0.04	36.75 \pm 0.17	87.93 \pm 0.03
+ ARPO-z′-V1	33.68 \pm 0.15	87.44 \pm 0.02	37.82 \pm 0.35	87.91 \pm 0.05
+ CLewR†	35.49 \pm 0.04	87.77 \pm 0.01	39.05 \pm 0.07	88.22 \pm 0.01
+ CLewR-z†	35.34 \pm 0.05	87.77 \pm 0.03	38.02 \pm 0.11	88.15 \pm 0.02
+ ARPO-z′-V2	33.96 \pm 0.09	87.49 \pm 0.10	37.47 \pm 0.15	87.84 \pm 0.06
+ CLewR†	35.53 \pm 0.03	87.77 \pm 0.04	38.89 \pm 0.07	88.17 \pm 0.03
+ CLewR-z†	35.43 \pm 0.01	87.80 \pm 0.04	37.87 \pm 0.09	88.13 \pm 0.04

Table 1: Results across several base models and preference optimization methods, with and without curriculum learning. Best overall results and per-model best results are shown in bold. A dagger (†) indicates statistical significance over the non-curriculum PO method.

sidering both en→zh and zh→en directions.

To evaluate the models, we compute BLEU and COMET-22 across all considered languages for both en→xx and xx→en directions. For each setup, we perform three training runs and report the average scores and the corresponding standard deviations across runs. For each run, we select the

checkpoint with the best validation performance.

To complement BLEU and COMET, we report additional metrics in the results. Specifically, we use the corpus-level Moving Average Type-Token Ratio (MATTR) (Covington and McFall, 2010) to measure lexical diversity, as well as TER (Snover et al., 2006), chrF (Popovic, 2015), and an LLM-as-a-judge metric (Zheng et al., 2023) to assess translation quality. TER measures edit distance, while chrF captures character n-gram overlap, being more robust to morphological variation. The LLM-as-a-judge prompt is given in Appendix A.3.

4 Results

We focus on two main comparisons: CLewR vs. CurriDPO, and CLewR vs. non-curriculum. In Table 1, we present MT results with various PO methods, with and without curriculum, applied on generic LLMs (Gemma2, Qwen2.5 and Llama3.1). **CLewR vs. CurriDPO.** Across all generic LLMs, DPOP+CLewR outperforms DPOP+CurriDPO in terms of COMET. The improvements, as measured via paired bootstrap resampling with 10,000 samples and a 95% confidence interval, are statistically significant on 2 out of the 3 models (Gemma2 and Qwen2.5). However, CLewR does not boost the performance of the standalone DPOP. We hypothesize that this is due to the DPOP margin, which depends on both policy π_θ and reference model $\pi_{\theta_{\text{ref}}}$: $m = m(\pi_\theta) - m(\pi_{\theta_{\text{ref}}})$. Because the reference margin is typically positive, it offsets part of the policy margin. As a result, the optimization signal in DPOP is weaker compared to other preference optimization methods that rely only on the policy margin. Indeed, CLewR brings consistent performance gains to highly competitive preference optimization methods, namely CPO and ARPO, which rely on a different *modus operandi*.

CLewR vs. base PO methods. CLewR consistently improves both BLEU and COMET scores across LLMs and PO algorithms. Most improvements are statistically significant according to paired bootstrap resampling with a 95% confidence interval over all sentence pairs. Moreover, our ARPO-z′-V1 and ARPO-z′-V2 variants further improve performance over the other preference methods. This indicates that by incorporating external semantic signals derived from MT metrics into z′, our versions of ARPO better adjust preference margins. ARPO-z′-V2 obtains the best performance on average. Additional experiments exploring more z′ combinations are presented in Appendix A.1, for

Model	Test on 3 languages (es, it, pt)			
	en→xx		xx→en	
	BLEU ↑	COMET ↑	BLEU ↑	COMET ↑
X-ALMA (ARPO)	37.50	89.00	39.40	88.80
Gemma-X2-9B	37.27	88.85	41.06	88.87
+ DPOP	38.23±0.01	89.01±0.02	41.93 ±0.10	89.09±0.02
+ CurriDPO	38.01±0.01	89.04±0.02	41.60±0.05	89.08±0.00
+ CLewR	38.24±0.16	88.93±0.01	41.80±0.07	89.09±0.02
+ CPO	36.36±0.04	89.04±0.00	41.00±0.09	89.05±0.01
+ CLewR†	37.69±0.03	88.87±0.01	41.45±0.14	89.04±0.01
+ ARPO	38.50±0.03	89.14 ±0.01	40.45±0.02	89.06±0.00
+ CLewR†	37.83±0.01	88.93±0.01	41.46±0.02	89.06±0.00
+ CLewR-z	38.35±0.01	89.12±0.00	40.29±0.02	89.02±0.01
+ ARPO-z'-V1	38.56±0.04	89.07±0.01	41.31±0.13	89.04±0.02
+ CLewR	38.08±0.01	88.98±0.00	41.72±0.02	89.09±0.00
+ CLewR-z	38.26±0.01	89.09±0.01	40.87±0.22	89.02±0.01
+ ARPO-z'-V2	38.58 ±0.04	89.07±0.00	41.33±0.02	89.04±0.00
+ CLewR	38.06±0.02	88.99±0.01	41.83±0.01	89.10 ±0.01
+ CLewR-z	38.24±0.03	89.08±0.01	40.91±0.25	89.02±0.03

Table 2: Results of GemmaX2 combined with preference optimization methods, with and without curriculum. The new variations of GemmaX2 surpass both the original model GemmaX2 and X-ALMA. The best results for BLEU and COMET across en→xx and xx→en directions are in bold. A dagger (†) indicates statistical significance over the non-curriculum PO method.

Model	Test on English↔Chinese			
	zh→en		en→zh	
	BLEU ↑	COMET ↑	BLEU ↑	COMET ↑
Gemma2-9B	27.84	86.90	38.27	81.18
+ ARPO-z'-V1	27.66±0.07	87.15±0.07	41.34±0.26	88.15±0.06
+ CLewR-z	29.49 ±0.02	87.43 ±0.04	42.46 ±0.10	88.22 ±0.02
+ ARPO-z'-V2	27.94±0.26	87.20±0.03	41.14±0.13	88.12±0.06
+ CLewR-z	29.15 ±0.35	87.40 ±0.06	42.75 ±0.14	88.20 ±0.06

Table 3: English↔Chinese results of Gemma2-9B based on ARPO-z', with and without curriculum. The results are averaged over three runs.

en→ro translations.

MT-adapted LLM. In Table 2, we compare the current state-of-the-art LLM for MT, GemmaX2, with PO-based variants and observe that the current SOTA can be further improved by applying preference optimization or curriculum learning. However, these gains are less pronounced than those previously observed in Table 1, and the strongest overall performance for GemmaX2 is obtained using DPOP. This suggests that LLMs optimized for MT may benefit less from more complex preference optimization methods and curriculum learning.

Chinese results. In Table 3, we report further experiments for en→zh and zh→en language pairs. Our curriculum strategy always improves both ARPO-z' variants on both directions and metrics.

Diversity metrics. We further report additional results in terms of diversity and translation quality

Model	MATTR ↑		TER ↓		chrF ↑	
	en→xx	xx→en	en→xx	xx→en	en→xx	xx→en
	Gemma2-9B	85.87	83.15	91.29	61.31	39.11
+ DPO	69.10	78.97	102.90	61.82	54.48	62.62
+ curriDPO	67.70	71.92	107.03	82.23	54.44	60.56
+ CLewR	68.62	76.96	103.15	68.61	54.38	61.98
+ CPO	83.15	82.64	54.36	51.20	60.07	63.04
+ CLewR	83.14	82.60	50.00	46.09	62.21	65.61
+ ARPO	83.06	82.75	51.02	49.18	61.49	64.16
+ CLewR	83.33	82.65	49.43	45.88	62.50	65.73
+ CLewR-z	83.53	82.75	49.81	46.74	62.41	65.74
+ ARPO-z'-V1	82.41	82.34	50.59	46.37	62.25	65.71
+ CLewR	83.18	82.79	48.70	45.03	63.00	66.43
+ CLewR-z	82.68	82.46	49.52	45.38	62.67	66.47
+ ARPO-z'-V2	82.56	82.41	50.34	46.78	62.37	65.54
+ CLewR	83.18	82.82	48.67	45.21	63.05	66.42
+ CLewR-z	82.77	82.38	49.34	45.18	62.74	66.61

Table 4: Diversity (MATTR) and performance (TER, chrF) metrics for Gemma2-9B. For each PO method, the best score is in bold.

metrics in Tables 4 and 7, respectively. Consistent with the trends observed for BLEU and COMET, CLewR generally provides improvements for the more advanced preference optimization methods in terms of lexical diversity (MATTR), translation quality (chrF), while also reducing edit distance as measured by TER. This indicates that the generated translations require fewer post-edits to match the reference, reflecting better overall adequacy and fluency alignment. We additionally analyze Borda count rankings in Appendix A.4, where we provide a more detailed breakdown across individual language pairs and translation directions.

LLM-as-a-Judge. We report LLM-as-a-judge evaluations in Appendix A.3.

Ablation study. We discuss ablation experiments in Appendix A.1.

5 Conclusion

In this paper, we introduced a novel curriculum learning strategy for preference optimization in MT. We demonstrated consistent gains across several general LLM families. We also showed that the adaptive penalty in ARPO based on modified distances z' derived from MT metrics yields additional performance gains. Moreover, we found that CLewR-z can further improve performance over the original curriculum scoring function.

In future work, we aim to explore the application of CLewR to other NLP tasks.

Acknowledgments

This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS - UEFISCDI, project number PN-IV-P1-PCE-2023-0354, within PNCDI IV.

Limitations

Due to computational constraints, we limited our experiments to the 6 languages in the Romance language group as defined in X-ALMA. Generalization is demonstrated only on Chinese. More experiments on other language groups and other model families would help strengthen the results.

Similar to generic PO algorithms, CLewR does not specifically address training data and model-specific biases. To mitigate potential biases, additional debiasing techniques need to be applied.

References

- Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. [Tower: An Open Multilingual Large Language Model for Translation-Related Tasks](#). In *Proceedings of Conference on Language Modeling*.
- Zachary Ankner, Naomi Saphra, Davis Blalock, Jonathan Frankle, and Matthew Leavitt. 2024. [Dynamic masking rate schedules for MLM pretraining](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 477–487.
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.
- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. [Curriculum learning](#). In *Proceedings of the International Conference on Machine Learning*, pages 41–48.
- Ernie Chang, Hui-Syuan Yeh, and Vera Demberg. 2021. [Does the Order of Training Samples Matter? Improving Neural Data-to-Text Generation with Curriculum Learning](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 727–733.
- Marta R. Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *arXiv preprint arXiv:2207.04672*.
- Michael A. Covington and Joe D. McFall. 2010. [Cutting the Gordian knot: The moving-average type–token ratio \(MATTR\)](#). *Journal of Quantitative Linguistics*, 17(2):94–100.
- Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, Nicu Sebe, and Mubarak Shah. 2025a. [Curriculum direct preference optimization for diffusion and consistency models](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2824–2834.
- Florinel-Alin Croitoru, Nicolae-Cătălin Ristea, Radu Tudor Ionescu, and Nicu Sebe. 2025b. [Learning rate curriculum](#). *International Journal of Computer Vision*, 133(1):291–314.
- Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. [Multilingual Machine Translation with Open Large Language Models at Practical Scale: An Empirical Study](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5420–5443.
- Peter Emerson. 2013. [The original Borda count and partial voting](#). *Social Choice and Welfare*, 40(2):353–358.
- Carlos Florensa, David Held, Markus Wulfmeier, Michael Zhang, and Pieter Abbeel. 2017. [Reverse curriculum generation for reinforcement learning](#). In *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78, pages 482–495.
- Yantao Gong, Cao Liu, Jiazhen Yuan, Fan Yang, Xunliang Cai, Guanglu Wan, Jiansong Chen, Ruiyao Niu, and Houfeng Wang. 2021. [Density-based dynamic curriculum learning for intent detection](#). In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3034–3037.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. [The Llama 3 Herd of Models](#). *arXiv preprint arXiv:2407.21783*.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-Rank Adaptation of Large Language Models](#). In *Proceedings of the International Conference on Learning Representations*.
- Andrei Jarca, Florinel-Alin Croitoru, and Radu Tudor Ionescu. 2024. [CBM: Curriculum by Masking](#). In *Proceedings of the 27th European Conference on Artificial Intelligence*, volume 392, pages 314–321.
- Andrei Jarca, Florinel-Alin Croitoru, and Radu Tudor Ionescu. 2025. [Task-informed anti-curriculum by masking improves downstream performance on text](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2192–2201.
- Diederik P. Kingma and Jimmy Lei Ba. 2015. [Adam: A method for stochastic gradient descent](#). In *Proceedings of the International Conference on Learning Representations*.

- James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. [Overcoming catastrophic forgetting in neural networks](#). *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.
- Tom Kocmi and Ondřej Bojar. 2017. [Curriculum Learning and Minibatch Bucketing in Neural Machine Translation](#). In *Proceedings of Recent Advances in Natural Language Processing*, pages 379–386.
- Cao Liu, Shizhu He, Kang Liu, and Jun Zhao. 2018. [Curriculum Learning for Natural Answer Generation](#). In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 4223–4229.
- Fengbei Liu, Yu Tian, Yuanhong Chen, Yuyuan Liu, Vasileios Belagiannis, and Gustavo Carneiro. 2022. [ACPL: Anti-curriculum Pseudo-labelling for Semi-supervised Medical Image Classification](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20706.
- Jinglin Liu, Yi Ren, Xu Tan, Chen Zhang, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2020a. [Task-level curriculum learning for non-autoregressive neural machine translation](#). In *Proceedings of the Twenty-Ninth International Joint Conferences on Artificial Intelligence*, pages 3861–3867.
- Xuebo Liu, Houtim Lai, Derek F. Wong, and Lidia S. Chao. 2020b. [Norm-based curriculum learning for neural machine translation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 427–436.
- Neelu Madan, Nicolae-Cătălin Ristea, Kamal Nasrollahi, Thomas B. Moeslund, and Radu Tudor Ionescu. 2024. [CL-MAE: Curriculum-Learned Masked Autoencoders](#). In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2492–2502.
- Koichi Nagatsuka, Clifford Broni-Bediako, and Masayasu Atsumi. 2023. [Length-Based Curriculum Learning for Efficient Pre-training of Language Models](#). *New Generation Computing*, 41(1):109–134.
- Sanmit Narvekar, Jivko Sinapov, Matteo Leonetti, and Peter Stone. 2016. [Source task creation for curriculum learning](#). In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 566–574.
- Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddhartha Naidu, and Colin White. 2024. [Smaug: Fixing failure modes of preference optimisation with DPO-positive](#). *arXiv preprint arXiv:2402.13228*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [BLEU: A method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.
- Pulkit Patnaik, Rishabh Maheshwary, Kelechi Ogueji, Vikas Yadav, and Sathwik Tejaswi Madhusudhan. 2024. [Enhancing alignment using curriculum learning & ranked preferences](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12891–12907.
- Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. [Competence-based curriculum learning for neural machine translation](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1162–1172.
- Eduard Gabriel Poesina, Cornelia Caragea, and Radu Tudor Ionescu. 2024. [A Novel Cartography-Based Curriculum Learning Method Applied on RoNLI: The First Romanian Natural Language Inference Corpus](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 236–253.
- Maja Popovic. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.
- Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2505.09388*.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. [Direct Preference Optimization: Your language model is secretly a reward model](#). In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, volume 36, pages 53728–53741.
- Ricardo Rei, José G. C. De Souza, Duarte Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André Martins. 2022. [COMET-22: Unbabel-IST 2022 submission for the metrics shared task](#). In *Proceedings of the Seventh Conference on Machine Translation*, pages 578–585.
- Samarth Sinha, Animesh Garg, and Hugo Larochelle. 2020. [Curriculum by smoothing](#). In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 21653–21664.
- Matthew G. Snover, Bonnie J. Dorr, Richard M. Schwartz, Linnea Micciulla, and John Makhoul. 2006. [A study of translation edit rate with targeted human annotation](#). In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2022. [Curriculum learning: A survey](#). *International Journal of Computer Vision*, 130:1526–1565.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. [Gemma 2: Improving open language models at a practical size](#). *arXiv preprint arXiv:2408.00118*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, and 1 others. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *arXiv preprint arXiv:2307.09288*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024a. [A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models](#). In *Proceedings of the International Conference on Learning Representations*.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2025. [X-ALMA: Plug & Play Modules and Adaptive Rejection for Quality Translation at Scale](#). In *Proceedings of the International Conference on Learning Representations*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024b. [Contrastive Preference Optimization: Pushing the boundaries of LLM performance in machine translation](#). In *Proceedings of the 41st International Conference on Machine Learning*, pages 55204–55224.

Runzhe Zhan, Xuebo Liu, Derek F Wong, and Lidia S Chao. 2021. [Meta-Curriculum Learning for Domain Adaptation in Neural Machine Translation](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14310–14318.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena](#). In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 46595–46623.

A Appendix

A.1 Ablation Study

In Table 5, we evaluate several z' variants and hyperparameters (η , β). The hyperparameter configuration for every evaluated variant is given in Table 6. The best variants of ARPO typically involve z_{COMET} . z_{BLEU} helps performance when averaged

Model	Test on Romanian	
	en→ro BLEU ↑	COMET ↑
Gemma2-9B	23.25	68.59
+ ARPO	38.18 \pm 0.55	90.54 \pm 0.02
+ CLewR	38.53 \pm 0.30	90.58 \pm 0.01
+ CLewR- z	38.49 \pm 0.20	90.49 \pm 0.05
+ ARPO $\eta = 0.1$	38.29 \pm 0.03	90.63 \pm 0.01
+ CLewR	39.78 \pm 0.02	90.79 \pm 0.01
+ CLewR- z	39.10 \pm 0.03	90.61 \pm 0.01
+ ARPO $\beta = 0.5$	35.62 \pm 0.10	90.26 \pm 0.03
+ CLewR	37.31 \pm 0.01	90.08 \pm 0.01
+ CLewR- z	36.96 \pm 0.07	90.41 \pm 0.01
+ ARPO- z' -V3	36.98 \pm 0.22	90.23 \pm 0.02
+ CLewR	38.43 \pm 0.01	90.32 \pm 0.01
+ CLewR- z	38.58 \pm 0.01	90.53 \pm 0.01
+ ARPO- z' -V4	37.29 \pm 0.14	90.34 \pm 0.05
+ CLewR	38.44 \pm 0.05	90.46 \pm 0.02
+ CLewR- z	38.46 \pm 0.04	90.39 \pm 0.01
+ ARPO- z' -V1	38.33 \pm 0.26	90.59 \pm 0.00
+ CLewR	39.71 \pm 0.06	90.75 \pm 0.02
+ CLewR- z	39.38 \pm 0.02	90.73 \pm 0.04
+ ARPO ARPO- z' -V5	37.31 \pm 0.00	90.18 \pm 0.01
+ CLewR	38.99 \pm 0.02	90.54 \pm 0.02
+ CLewR- z	38.49 \pm 0.07	90.43 \pm 0.02
+ ARPO- z' -V6	37.18 \pm 0.03	90.18 \pm 0.00
+ CLewR	38.64 \pm 0.03	90.52 \pm 0.01
+ CLewR- z	38.53 \pm 0.01	90.46 \pm 0.02
+ ARPO- z' -V7	38.06 \pm 0.03	90.58 \pm 0.01
+ CLewR	39.52 \pm 0.04	90.63 \pm 0.01
+ CLewR- z	38.70 \pm 0.01	90.77 \pm 0.01
+ ARPO- z' -V8	38.23 \pm 0.01	90.51 \pm 0.02
+ CLewR	39.59 \pm 0.07	90.76 \pm 0.05
+ CLewR- z	39.62 \pm 0.04	90.86 \pm 0.01
+ ARPO- z' -V9	36.79 \pm 0.08	90.30 \pm 0.09
+ CLewR	38.43 \pm 0.06	90.41 \pm 0.05
+ CLewR- z	37.98 \pm 0.15	90.31 \pm 0.07
+ ARPO- z' -V2	38.16 \pm 0.11	90.50 \pm 0.02
+ CLewR	39.66 \pm 0.09	90.72 \pm 0.01
+ CLewR- z	39.58 \pm 0.03	90.86 \pm 0.01

Table 5: Ablation study on translations from English to Romanian, experimenting with different hyperparameters and variations of z for ARPO.

with z_{θ} and z_{COMET} . Smaller values for η_1 , η_2 and η_3 lead to better performance.

A.2 Overview of Performance Gains

In Figure 1, we present a detailed analysis on each configuration and its statistical improvements. For each model, the most significant increase is obtained by applying CLewR over CPO. CLewR over DPO degrades performance, suggesting that our curriculum is more beneficial for advanced preference algorithms.

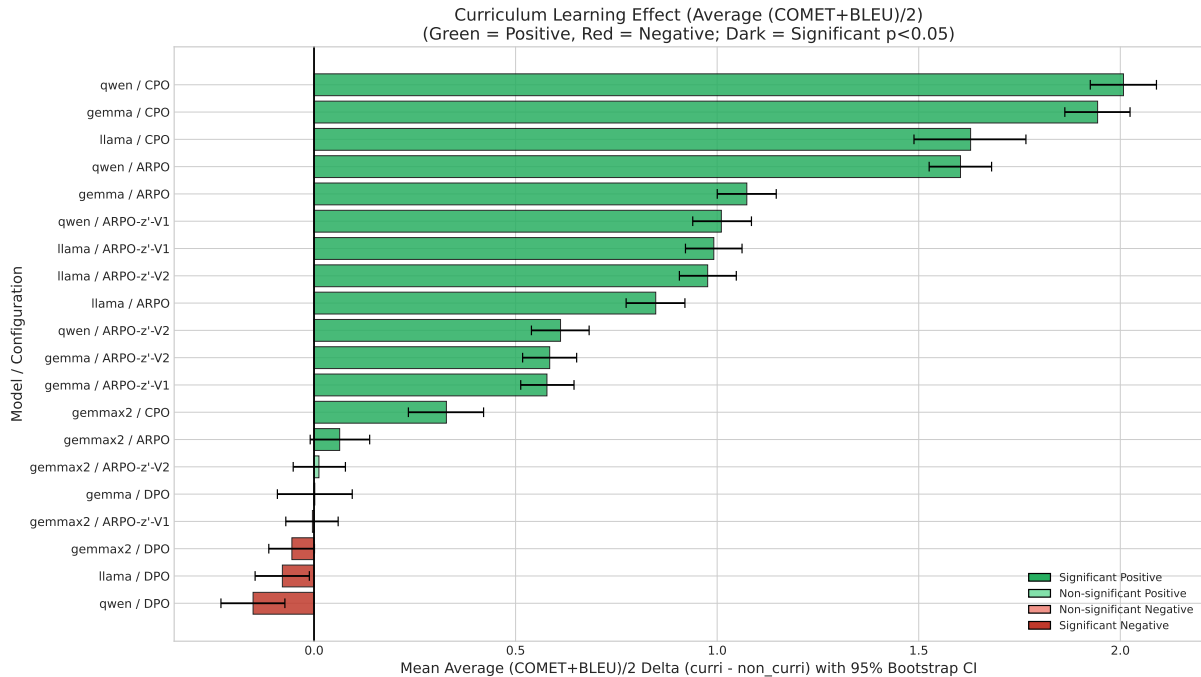


Figure 1: Tested configurations and their improvements from curriculum against the non-curriculum variant. 13 out of 20 configurations show statistically significant improvements, 4 show no significant change, and 3 result in performance degradation. The black bars represent 95% confidence intervals. Best viewed in color.

ARPO Version	η_1	η_2	η_2
ARPO	1.5	0	0
ARPO- z' -V1	0	0	0.5
ARPO- z' -V2	$\frac{0.1}{3}$	$\frac{0.1}{3}$	$\frac{0.5}{3}$
ARPO- z' -V3	0	1.5	0
ARPO- z' -V4	0	0	6
ARPO- z' -V5	0	$\frac{1.5}{2}$	$\frac{6}{2}$
ARPO- z' -V6	$\frac{1.5}{2}$	$\frac{1.5}{2}$	0
ARPO- z' -V7	$\frac{1.5}{2}$	0	$\frac{6}{2}$
ARPO- z' -V8	$\frac{0.1}{2}$	0	$\frac{0.5}{2}$
ARPO- z' -V9	$\frac{1.5}{3}$	$\frac{1.5}{3}$	$\frac{6}{3}$

Table 6: ARPO variants and their hyperparameter configurations.

A.3 LLM-as-a-Judge Results

In addition to automatic metrics, we evaluate translation quality using an LLM-as-a-judge setup, where a large language model assigns a scalar score to each translation based on adequacy and fluency. This complementary evaluation provides a more holistic assessment that is less dependent on surface-level overlap with reference translations.

We use GPT5.1 (reasoning effort “medium”) to judge MT results, with the following prompt:

Model	LLM-as-a-judge	
	en→xx	xx→en
Gemma2-9B		
+ DPO	84.30	92.59
+ CLewR	83.71	92.49
+ CPO	85.40	92.08
+ CLewR	87.99	93.99
+ ARPO	88.84	94.15
+ CLewR	88.54	94.42
+ ARPO- z' -V1	90.23	94.76
+ CLewR	90.00	95.14
+ ARPO- z' -V2	90.41	94.80
+ CLewR	90.04	95.21

Table 7: LLM-as-a-judge evaluation results for Gemma2-9B. Results are averaged after running on the three predictions for each PO method.

JUDGE_PROMPT_TEMPLATE

Score the following translation from {source_lang_name} to {target_lang_name} on a scale from 0 to 100, where a score of 0 means a broken or poor translation; 33 indicates a flawed translation with significant issues; 66 indicates a good translation with only minor issues in grammar, fluency, or consistency; and 100 represents a perfect translation in both meaning and grammar. Answer with only a whole number.

{source_lang_name} source text: {source_text}
{target_lang_name} translation: {model_output}

The LLM-as-a-judge results are reported in Table 7. We observe that the GPT5.1 judge produces

Examples of training triplets: easy (low similarity between chosen and rejected) and hard (high similarity)

pt→en

EAS : “Para retornar ao nível anterior de energia, seria necessário descarregar a carga extra que foi recebida da iluminação.”

chosen: “To return to their previous energy level, they must get rid of the extra energy they got from the light.”

rejected: “To return to the previous level of energy, it would be necessary to discharge the extra charge that was received from the lighting.”

es→en

EAS : “Agaché la cabeza y las lágrimas me caían todo el tiempo.”

chosen: “I had my head down, and tears were falling the whole time.”

rejected: “My head was down and I was crying the whole time.”

it→en

HARD: “L’arte della ceramica richiede pazienza e precisione, con artigiani che modellano l’argilla con maestria e la trasformano in pezzi unici e preziosi, ciascuno con la propria storia e carattere.”

chosen: “The art of ceramics requires patience and precision, with artisans masterfully shaping the clay into unique and precious pieces, each with its own story and character.”

rejected: “The art of ceramics requires patience and precision, with artisans skillfully shaping the clay into unique and treasured pieces, each with its own story and character.”

es→en

HARD: “Las nuevas medidas sanitarias han ayudado a controlar la propagación de enfermedades en la comunidad, protegiendo la salud pública y mejorando la calidad de vida.”

chosen: “The new health measures have helped control the spread of diseases in the community, safeguard public health, and improve the quality of life.”

rejected: “The new health measures have helped control the spread of diseases in the community, protecting public health and improving the quality of life.”

Figure 2: Easy vs. hard examples for curriculum learning based on the score s derived from BLEU, METEOR and COMET metrics.

results that are mostly consistent with the usual BLEU and COMET performance metrics.

A.4 Detailed Results and Borda Counts

To provide a more comprehensive comparison across language pairs and translation directions, we report Borda count rankings (Emerson, 2013) based on BLEU and COMET scores. The detailed results are presented in Tables 8 and 9, which include per-language-pair scores, as well as aggregated rankings.

For each metric, systems are first ranked independently on each language pair and translation direction. We then compute the average rank across all language pairs, corresponding to a standard Borda count aggregation. We consider two variants of this procedure: (i) *per model* (M), where all model configurations are jointly ranked for each language pair, and (ii) *per method* (m), where each base method is ranked together with its corresponding variants, e.g. CLewR and CLewR- z .

Since the number of systems N differs between these settings, we linearly rescale the average rank to the interval $[0, 100]$, where 100 corresponds to a consistently top rank, and 0 to a consistently bottom rank across all language pairs. The rescaled

score is computed as:

$$100 \cdot \left(1 - \frac{R - 1}{N - 1}\right), \quad (5)$$

where R denotes the average rank. The last columns in Tables 8 and 9 report these normalized Borda scores for both BLEU and COMET, under the per-model (M) and per-method (m) settings, respectively.

Overall, the Borda rankings are consistent with the trends observed in the main experiments, confirming that CLewR-based variants generally achieve stronger aggregate performance across language pairs and directions. In particular, the improvements are reflected in both BLEU-based and COMET-based rankings under both per-model and per-method settings. This further supports the robustness of our conclusions beyond individual metric values.

A.5 Easy vs. Hard Examples for Curriculum

In Figure 2, we list several (x, y_w, y_l) triplets that are either easy (low similarity $s(y_w, y_l)$) or hard (high similarity $s(y_w, y_l)$). For the hard examples, we underline the differences between chosen and rejected sentences.

Model	en→ca		en→es		en→gl		en→it		en→pt		en→ro		Borda (M)		Borda (m)	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
	Gemma2-9B	12.35	56.90	15.62	57.20	18.57	62.96	19.15	64.46	37.66	76.35	15.74	56.63	7.14	0.00	
+ DPO	26.14	84.98	19.74	85.31	16.33	82.40	19.55	86.75	31.30	88.37	26.49	87.82	11.90	15.48	50.00	58.33
+ CurriDPO	26.41	84.57	16.02	84.56	17.81	82.93	20.28	86.52	27.35	87.77	23.00	87.46	13.10	9.52	58.33	16.67
+ CLewR	22.72	84.87	18.56	85.46	16.83	83.22	20.28	87.01	32.91	88.63	22.77	87.45	11.90	17.86	50.00	75.00
+ CPO	36.72	86.42	28.21	86.49	29.32	85.03	28.54	88.10	44.18	89.17	34.22	89.37	28.57	28.57	0.00	0.00
+ CLewR	40.33	86.78	28.33	86.61	34.59	86.14	29.58	88.11	47.47	89.63	37.17	89.83	48.81	35.71	100.00	100.00
+ ARPO	38.56	86.93	28.70	86.97	33.28	86.61	29.51	88.61	46.09	89.74	36.10	89.86	38.10	52.38	8.33	50.00
+ CLewR	40.59	87.21	28.64	86.83	35.02	86.52	29.91	88.33	47.96	89.84	37.66	89.97	61.90	51.19	58.33	41.67
+ CLewR-z	40.06	87.04	28.78	86.85	35.07	86.66	29.95	88.46	47.33	89.64	37.79	90.05	60.71	51.19	83.33	58.33
+ ARPO-z'-V1	38.91	87.16	29.66	87.09	34.49	87.20	30.29	88.68	47.52	89.94	37.21	90.32	61.90	83.33	16.67	58.33
+ CLewR	41.44	87.47	29.22	86.99	35.74	87.02	30.42	88.52	48.91	89.98	38.93	90.42	91.67	85.71	83.33	58.33
+ CLewR-z	40.80	87.28	28.73	86.99	35.20	87.17	30.43	88.72	48.49	89.89	38.22	90.24	78.57	78.57	50.00	41.67
+ ARPO-z'-V2	39.64	87.22	29.56	87.09	34.41	87.28	30.15	88.67	47.69	89.97	37.32	90.32	61.90	89.29	16.67	75.00
+ CLewR	41.45	87.34	29.14	87.01	35.72	86.97	30.40	88.50	48.97	89.97	38.99	90.42	91.67	82.14	83.33	58.33
+ CLewR-z	40.47	87.21	28.91	86.98	35.55	87.14	30.69	88.73	48.55	89.94	38.29	90.28	83.33	77.38	50.00	25.00
Qwen2.5-7B	3.26	44.52	3.50	47.39	2.98	43.89	3.19	45.88	6.57	48.85	5.11	45.13	0.00	0.00		
+ DPO	23.80	78.95	24.36	85.29	17.19	80.00	22.67	85.94	36.17	87.96	22.41	79.25	16.67	21.43	66.67	91.67
+ CurriDPO	24.99	78.19	22.14	84.29	18.53	78.72	22.35	85.25	32.30	87.10	22.50	77.83	16.67	7.14	66.67	0.00
+ CLewR	23.52	78.92	21.51	84.64	18.08	80.36	21.88	85.60	34.52	87.71	22.02	79.21	9.52	15.48	16.67	58.33
+ CPO	28.07	80.46	25.85	85.61	24.32	81.95	23.75	85.97	38.65	88.11	25.42	81.46	29.76	30.95	16.67	33.33
+ CLewR	31.51	81.15	25.80	85.31	27.37	82.49	25.14	85.70	42.01	88.28	28.50	81.96	35.71	32.14	83.33	66.67
+ ARPO	31.58	83.26	26.87	86.36	28.41	84.46	25.20	86.82	41.49	88.74	28.93	84.17	46.43	70.24	16.67	83.33
+ CLewR	33.76	82.81	26.53	85.99	29.14	83.84	25.99	86.76	43.61	88.79	30.33	84.07	65.48	46.43	75.00	25.00
+ CLewR-z	32.89	82.87	26.67	86.07	29.28	84.28	25.63	86.71	42.41	88.67	29.95	84.18	58.33	53.57	58.33	41.67
+ ARPO-z'-V1	32.94	83.04	27.60	86.40	28.64	84.26	25.68	86.98	42.21	88.90	29.37	84.46	64.29	75.00	16.67	33.33
+ CLewR	34.78	83.43	27.41	86.20	30.17	84.15	26.84	86.89	44.91	89.10	31.94	85.11	89.29	76.19	75.00	33.33
+ CLewR-z	34.61	83.64	27.18	86.34	30.61	84.35	26.72	87.11	44.81	89.11	32.16	85.01	84.52	91.67	58.33	83.33
+ ARPO-z'-V2	32.69	83.30	27.48	86.28	28.27	84.20	25.86	86.99	42.52	88.88	28.87	84.37	59.52	70.24	16.67	25.00
+ CLewR	34.99	83.52	27.42	86.21	30.12	84.10	26.76	86.78	44.85	89.03	32.09	85.07	89.29	71.43	75.00	33.33
+ CLewR-z	34.46	83.62	27.22	86.33	30.78	84.36	26.70	87.13	44.86	89.10	31.99	85.00	84.52	89.29	58.33	91.67
LLaMA3.1-8B	1.95	42.78	1.34	46.25	1.49	45.12	1.42	47.07	2.18	48.00	1.82	46.75	7.14	0.00		
+ DPO	34.06	86.08	27.28	85.83	27.16	85.08	27.36	87.66	40.96	88.76	32.33	88.87	28.57	36.90	50.00	83.33
+ CurriDPO	35.92	86.10	26.71	85.80	30.06	85.18	27.62	87.43	40.98	88.59	33.47	88.56	38.10	28.57	91.67	50.00
+ CLewR	33.36	85.93	26.48	85.62	27.82	85.03	27.04	87.46	39.02	88.45	31.51	88.57	22.62	23.81	8.33	16.67
+ CPO	1.11	52.39	1.23	60.19	1.31	50.18	2.32	61.93	1.63	56.53	1.23	48.62	1.19	8.33	0.00	16.67
+ CLewR	2.70	48.81	4.99	66.26	1.42	56.03	4.77	65.96	7.59	71.15	2.33	49.78	13.10	13.10	100.00	83.33
+ ARPO	36.61	86.40	27.45	86.46	32.68	86.30	27.62	87.82	43.15	89.03	33.47	88.83	47.62	55.95	16.67	58.33
+ CLewR	37.90	86.40	27.35	86.37	33.30	86.12	28.05	87.69	45.17	89.28	35.16	88.80	64.29	51.19	66.67	25.00
+ CLewR-z	37.72	86.60	27.49	86.55	33.36	86.26	27.53	87.81	44.53	89.13	35.32	89.07	60.71	67.86	66.67	75.00
+ ARPO-z'-V1	36.81	86.36	28.19	86.54	31.99	86.06	27.92	87.84	43.31	89.05	33.84	88.78	57.14	53.57	16.67	8.33
+ CLewR	38.95	86.86	27.90	86.53	34.80	86.49	28.52	88.00	46.16	89.38	36.64	89.34	86.90	86.90	66.67	58.33
+ CLewR-z	39.08	86.97	27.91	86.63	34.33	86.28	28.53	88.02	45.59	89.40	36.56	89.34	84.52	90.48	66.67	91.67
+ ARPO-z'-V2	37.18	86.45	28.43	86.50	32.54	86.15	28.00	87.84	43.64	89.07	33.99	88.93	64.29	61.90	16.67	0.00
+ CLewR	39.02	86.98	27.94	86.52	34.89	86.48	28.67	87.86	46.12	89.48	36.59	89.32	91.67	85.71	75.00	75.00
+ CLewR-z	39.31	87.05	28.08	86.70	34.29	86.37	28.50	87.98	45.89	89.40	36.49	89.27	84.52	90.48	58.33	75.00

Table 8: Results across several base models and preference optimization methods, with and without CLewR, for each pair of languages of the form en→xx. Normalized Borda counts are computed per model (M) and per method (m), respectively.

Model	ca→en		es→en		gl→en		it→en		pt→en		ro→en		Borda (M)		Borda (m)	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
	Gemma2-9B	37.48	84.90	26.41	83.38	34.35	85.40	30.35	86.63	44.79	87.32	36.52	85.79	21.43	0.00	
+ DPO	38.82	88.08	28.61	87.12	25.88	86.19	29.56	87.66	41.29	88.89	34.84	88.67	15.48	20.24	91.67	83.33
+ CurriDPO	33.84	87.60	25.10	86.73	22.57	85.62	27.14	87.34	32.52	88.17	28.84	88.18	1.19	8.33	8.33	0.00
+ CLewR	37.35	88.01	28.07	87.16	22.54	85.63	30.43	87.70	38.14	88.74	34.63	88.66	9.52	17.86	50.00	66.67
+ CPO	41.61	88.17	30.74	87.17	29.92	85.57	32.70	87.82	43.95	88.97	39.70	88.86	26.19	25.00	0.00	0.00
+ CLewR	45.53	88.82	32.76	87.63	36.95	87.15	35.03	88.31	48.24	89.58	43.52	89.46	59.52	48.81	100.00	100.00
+ ARPO	42.58	88.60	31.08	87.49	33.72	86.95	33.28	88.12	44.77	89.34	40.64	89.32	33.33	35.71	0.00	0.00
+ CLewR	45.55	88.84	32.95	87.67	37.17	87.42	35.02	88.35	48.48	89.61	43.46	89.49	61.90	59.52	91.67	83.33
+ CLewR-z	44.16	88.74	32.53	87.70	38.93	88.03	34.32	88.27	47.36	89.56	41.95	89.43	47.62	51.19	58.33	66.67
+ ARPO-z'-V1	44.86	88.83	33.23	87.72	37.58	87.71	35.42	88.38	47.77	89.59	43.10	89.49	60.71	65.48	0.00	0.00
+ CLewR	45.98	88.97	33.45	87.78	39.14	88.05	35.67	88.47	49.07	89.74	44.36	89.66	90.48	86.90	75.00	58.33
+ CLewR-z	45.84	89.04	33.80	87.82	39.59	88.31	35.70	88.50	48.77	89.76	43.64	89.64	85.71	95.24	75.00	91.67
+ ARPO-z'-V2	45.00	88.86	33.24	87.71	36.36	87.44	35.26	88.38	47.65	89.53	43.10	89.54	57.14	63.10	0.00	0.00
+ CLewR	45.93	88.93	33.47	87.79	39.03	88.05	35.48	88.44	48.92	89.72	43.70	89.59	83.33	80.95	50.00	50.00
+ CLewR-z	46.04	89.03	33.92	87.87	40.32	88.40	35.72	88.50	48.96	89.78	43.74	89.64	97.62	97.62	100.00	100.00
Qwen2.5-7B	10.50	63.35	7.67	62.83	10.90	64.15	8.30	62.67	11.54	63.22	8.66	60.76	0.00	0.00		
+ DPO	36.38	87.33	28.19	87.03	28.82	86.17	29.62	87.54	40.58	88.74	34.73	87.71	35.71	53.57	100.00	100.00
+ CurriDPO	30.94	86.60	25.34	86.47	26.04	85.43	26.44	87.03	35.83	88.12	30.14	87.00	8.33	20.24	8.33	8.33
+ CLewR	34.85	87.21	26.38	86.78	22.51	85.29	28.68	87.32	38.37	88.48	33.02	87.51	17.86	27.38	41.67	41.67
+ CPO	35.29	86.38	27.77	86.12	27.04	84.43	28.65	86.61	37.99	87.69	33.20	86.93	20.24	7.14	0.00	0.00
+ CLewR	39.69	87.24	30.82	86.91	33.37	85.95	32.04	87.45	43.80	88.56	37.95	87.84	70.24	40.48	100.00	100.00
+ ARPO	35.81	86.76	27.92	86.41	27.87	84.82	28.59	86.92	38.01	87.95	34.51	87.41	25.00	16.67	0.00	0.00
+ CLewR	40.56	87.62	31.61	87.22	32.72	86.24	32.47	87.65	44.12	88.94	38.70	88.04	79.76	84.52	100.00	100.00
+ CLewR-z	38.99	87.49	29.79	87.08	31.97	86.07	30.49	87.52	41.68	88.62	36.37	87.92	44.05	57.14	50.00	50.00
+ ARPO-z'-V1	38.95	87.42	30.25	86.99	32.01	86.00	30.70	87.38	41.99	88.56	36.84	87.89	48.81	46.43	0.00	8.33
+ CLewR	41.29	87.61	31.69	87.10	33.43	86.22	32.86	87.66	44.22	88.86	39.55	88.13	95.24	80.95	100.00	75.00
+ CLewR-z	40.94	87.71	31.18	87.22	32.91	85.97	31.91	87.51	44.02	88.87	38.30	88.11	76.19	76.19	50.00	66.67
+ ARPO-z'-V2	40.11	87.78	31.19	87.27	34.43	86.77	31.41	87.68	43.39	88.85	37.88	88.12	69.05	91.67	25.00	75.00
+ CLewR	41.37	87.71	31.74	87.17	33.17	86.21	33.05	87.74	44.15	88.86	39.50	88.16	94.05	88.10	91.67	66.67
+ CLewR-z	40.74	87.59	30.67	87.21	32.13	85.76	31.75	87.58	43.61	88.80	37.99	88.10	65.48	64.29	33.33	8.33
LLaMA3.1-8B	2.11	44.96	1.43	45.01	1.81	44.15	1.57	47.36	2.12	47.70	2.03	48.65	0.00	0.00		
+ DPO	40.90	88.20	30.22	87.18	32.05	86.32	31.73	87.83	42.21	88.73	37.54	88.64	32.14	46.43	50.00	58.33
+ CurriDPO	40.88	87.99	29.92	87.09	33.36	86.51	30.52	87.68	43.46	88.66	37.92	88.56	34.52	28.57	58.33	16.67
+ CLewR	39.33	88.17	30.19	87.19	31.25	86.25	31.93	87.84	42.46	88.83	37.60	88.68	30.95	51.19	41.67	75.00
+ CPO	13.01	74.51	10.33	75.97	6.30	70.18	10.74	76.56	13.03	73.82	8.73	74.85	7.14	7.14	0.00	0.00
+ CLewR	15.39	77.11	11.54	77.73	10.90	72.15	12.27	77.68	15.14	77.67	13.19	78.24	14.29	14.29	100.00	100.00
+ ARPO	39.72	87.95	30.06	86.99	33.93	86.75	31.44	87.57	42.43	88.68	37.94	88.57	34.52	29.76	0.00	8.33
+ CLewR	42.50	88.26	31.58	87.24	34.70	86.62	33.20	87.77	45.54	88.97	40.42	88.82	78.57	65.48	100.00	83.33
+ CLewR-z	40.98	88.22	30.69	87.20	34.30	86.78	32.21	87.72	43.43	88.86	38.85	88.78	50.00	58.33	50.00	58.33
+ ARPO-z'-V1	42.01	88.13	31.60	87.17	34.73	86.89	33.16	87.74	44.60	88.77	40.85	88.75	71.43	48.81	16.67	0.00
+ CLewR	43.67	88.54	32.06	87.38	36.34	87.29	34.18	87.94	46.33	89.17	41.71	89.01	98.81	96.43	100.00	91.67
+ CLewR-z	42.06	88.38	31.37	87.32	35.55	87.08	32.88	87.93	45.41	89.10	40.86	89.09	72.62	85.71	33.33	58.33
+ ARPO-z'-V2	42.12	88.15	31.45	87.16	33.14	86.36	33.06	87.73	44.65	88.81	40.41	88.82	63.10	46.43	25.00	0.00
+ CLewR	43.30	88.52	32.42	87.34	35.61	87.11	34.16	87.94	46.18	89.13	41.66	89.01	92.86	91.67	91.67	83.33
+ CLewR-z	42.21	88.39	31.07	87.25	36.07	87.33	32.68	87.81	44.89	88.99	40.28	89.02	69.05	83.33	33.33	66.67

Table 9: Results across several base models and preference optimization methods, with and without CLewR, for each pair of languages of the form $xx \rightarrow en$. Normalized Borda counts are computed per model (M) and per method (m), respectively.