Modeling User Preferences with Automatic Metrics: Creating a High-Quality Preference Dataset for Machine Translation

Anonymous ACL submission

Abstract

Alignment with human preferences is an important step in developing accurate and safe large language models. This is no exception in machine translation (MT), where better handling of language nuances and context-specific variations leads to improved quality. However, preference data based on human feedback can be very expensive to obtain and curate at a large scale. Automatic metrics, on the other hand, can induce preferences, but they might not match human expectations perfectly. In this paper, we propose an approach that leverages the best of both worlds. We first collect sentence-level quality assessments from professional linguists on translations generated by multiple high-quality MT systems and evaluate the ability of current automatic metrics to recover these preferences. We then use this analysis to curate a new dataset, MT-PREF (**M**etric-induced **T**ranslation **PREF**erence), which comprises 18k instances covering 18 language directions, using texts sourced from multiple domains post-2022. We show that aligning TOWER models on MT-PREF significantly improves translation quality on WMT23 and FLORES benchmarks.[1]

1 Introduction

The use of large language models (LLMs) in machine translation (MT) has garnered significant attention from the research community (Kocmi et al., 2023). Unlike traditional sequence-to-sequence MT models trained on parallel data (Koehn and Knowles, 2017), LLM-based MT systems either use in-context learning to elicit translation knowledge acquired during pre-training (Briakou et al., 2023) or undergo supervised finetuning (SFT) on high-quality translations to further enhance their translation capabilities (Li et al., 2024; Xu et al., 2023; Alves et al., 2023, 2024).

The default SFT approach for LLM-based MT is to tune systems based solely on *single* human reference translations. However, this kind of supervision might be insufficient to push quality further: First, because *many* valid translations may exist for a given source, with some *preferred* over others (Mayhew et al., 2020). Second, because the next-token prediction objective of SFT does not capture sentence-level semantics and quality criteria (Eikema and Aziz, 2020; Liu et al., 2022). This has motivated new approaches which go beyond SFT to leverage translation preferences or quality feedback to improve learning (Yang et al., 2023; He et al., 2024; Xu et al., 2024; Zhu et al., 2024).

A key factor in aligning LLMs toward translation preferences is ensuring the quality and diversity of the datasets used for training (Gao et al., 2024; Morimura et al., 2024; Liu et al., 2023). Unfortunately, existing datasets have several limitations: First, they are created from translation outputs of one or two models, for limited language pairs, thereby restricting their diversity and applicability to novel scenarios. Second, these datasets are either entirely automatically generated (Xu et al., 2024) or completely based on human feedback (Zhu et al., 2024). While automatic evaluation offers efficiency, it lacks the crucial validation that the metrics used truly align with human preferences. On the other hand, datasets that use human feedback, while high-quality and reliable, pose resource constraints and are challenging to scale.

To bridge this gap, we provide a holistic approach to balance the advantages of automated metrics while ensuring that they lead to preferences that truly align with humans. We first collect sentence-level quality assessments and preferences from human expert translators (§3)—we use the WMT23 English-German and Chinese-English datasets (Kocmi et al., 2023) with outputs from five high-quality MT systems: TOWERINSTRUCT-7B, TOWERINSTRUCT-13B (Alves et al., 2024);

[1] We will release the code and datasets to reproduce all the results on acceptance.

ALMA-13B-R (Xu et al., 2024); GPT-4-based (Hendy et al., 2023) and GOOGLETRANSLATE. Using these assessments, we then examine the ability of automatic quality estimation (QE) metrics to recover human preferences. Our findings show that an ensemble of XCOMET-XL and XCOMET-XXL (Guerreiro et al., 2023)—XCOMET-XL+XXL—achieves the highest correlation with human judgments and a high precision score in identifying the preferred translations.

Using this analysis, we create a new MT preference dataset, MT-PREF (**M**etric-induced **T**ranslation **PREF**erence dataset), with source sentences mined post-2022 for 10 languages (English, German, Chinese, Russian, Portuguese, Italian, French, Spanish, Korean and Dutch). Translations for each source sentence are generated using diverse MT systems representing different architectures, training data, and quality levels (§4). We use the ensemble metric XCOMET-XL+XXL to get the most and least preferred translations from the set of hypotheses. Experiments on aligning MT-specialized decoder-only models (TOWER) using existing preference learning algorithms with our MT-PREF dataset demonstrate improved translation quality on the WMT23 (Kocmi et al., 2023) and FLORES (Costa-jussà et al., 2022) benchmarks, with larger gains in out-of-English translation directions (§6). Further analysis shows that the aligned models better rank translations according to human preferences over baselines.

2 Background: Aligning MT with Translation Preferences

Given a source text, the goal of MT is to generate a translation that accurately reflects the information and meaning conveyed in the source. At training time, the MT model π_θ goes through SFT to minimize the negative log-likelihood (NLL) loss induced by source-reference pairs (x, y):

$$\mathcal{L}_{\text{NLL}}(x, y; \theta) = -\log \pi_\theta\left(y|x\right). \tag{1}$$

A drawback of SFT is that it typically optimizes the model towards a *single* reference translation. In contrast, preference learning objectives incorporate relative preferences between alternatives, allowing the model to learn from subtle differences in translation quality (Zeng et al., 2023).

Different variants of preference optimization (PO) have been proposed in the literature. Reinforcement learning from human feedback (RLHF) has shown to be effective in aligning model behavior with human values (Ouyang et al., 2022). Rafailov et al. (2024) propose direct preference optimization (DPO) as a simple and scalable alternative to RLHF. Given a preference dataset $\mathcal{D}$ with source sentences x, preferred or chosen outputs y_+ and less preferred or rejected outputs y_-, the model is trained with the following objective:

$$\mathcal{L}_{\text{DPO}}(x, y_\pm; \pi_\theta, \pi_{\text{ref}}) = \tag{2}$$
$$-\log \sigma\left(\beta \log \frac{\pi_\theta(y_+|x)}{\pi_{\text{ref}}(y_+|x)} - \beta \log \frac{\pi_\theta(y_-|x)}{\pi_{\text{ref}}(y_-|x)}\right),$$

where π_θ is the parameterized policy, π_{ref} is a base reference policy (set to the policy used to generate the dataset for collecting preferences), and β is a (inverse) temperature hyperparameter.

One notable limitation of the DPO objective is that it requires both π_θ and π_{ref} in memory, significantly increasing memory requirements and computation costs. To address this, Xu et al. (2024) further approximate the DPO objective using a uniform reference model ($\pi_{\text{ref}} = \mathcal{U}$) to derive a contrastive preference optimization (CPO) loss:

$$\mathcal{L}_{\text{DPO}}(x, y_\pm; \pi_\theta, \mathcal{U}) = -\log \sigma\left(\beta \log \frac{\pi_\theta(y_+|x)}{\pi_\theta(y_-|x)}\right). \tag{3}$$

However, both losses (2)–(3) only maximize the relative difference between preferred and dispreferred outputs. On tasks like MT where the difference in the two outputs is small, this may lead to failure modes where the learning objective leads to a reduction of the model's likelihood of the preferred examples, as long as the relative probability between the two classes increases (Pal et al., 2024). Therefore, following Hejna et al. (2023), Xu et al. (2024) introduce a behavior cloning regularizer to ensure that the model stays close to the preferred distribution, leading to the final CPO objective:

$$\mathcal{L}_{\text{CPO}}(x, y_\pm; \theta) = \tag{4}$$
$$\mathcal{L}_{\text{DPO}}(x, y_\pm; \pi_\theta, \mathcal{U}) + \lambda \mathcal{L}_{\text{NLL}}(x, y_+; \theta),$$

where λ is a hyperparameter that controls the relative strength of the two objectives.

As the quality of the preference datasets used for training is key for its success (Gao et al., 2024; Morimura et al., 2024; Liu et al., 2023), we next discuss our process of collecting a high-quality dataset for preference learning for MT.

3 Modeling User Preferences Via Automatic Metrics

To create a high-quality preference dataset for MT, we need human judgments on translation outputs from strong MT systems. This helps us understand and model human preferences among competitive translations. Since large-scale collection of these judgments is costly, we evaluate existing automatic metrics to see if they effectively reflect human preferences. This determines if metrics can be reliable proxies for human judgments when translation quality is high and preference variance is low.

We describe the dataset, models, and task instructions given to the expert annotators used in our study in §3.1. The human evaluation results are presented in §3.2. Finally, we discuss our meta-evaluation of automatic MT metrics in their ability to recover human preferences in §3.3.

3.1 Data and Annotation Task

We randomly sample 200 source instances from the WMT23 English-German (EN-DE) and Chinese-English (ZH-EN) test sets and generate translations using five MT models: GOOGLETRANSLATE, GPT-4, TOWERINSTRUCT-7B, TOWERINSTRUCT-13B, and ALMA-13B-R (described in Appendix B).[2] We employ DA+SQM (Direct Assessment + Scalar Quality Metric) source contrastive evaluation (Kocmi et al., 2022), using the Appraise evaluation framework (Federmann, 2018).[3] We then ask one linguist per language pair to read all translations for a given source and evaluate each of them on a continuous 0-100 scale. The scale features seven labeled tick marks indicating different quality labels combining *accuracy* and *grammatical correctness*. Linguists can further adjust their ratings to reflect preferences or assign the same score to translations of similar quality. Detailed guidelines and a screenshot of the interface are provided in Appendix A. This results in a preference dataset including 1000 ratings each for EN-DE and ZH-EN.[4]

3.2 Human Evaluation Findings

We present the results from our human evaluation in Table 1 and discuss the findings below:

[2]This is the only dataset that was not used in the training of any evaluated models.

[3]https://github.com/AppraiseDev/Appraise.git.

[4]Completing the task takes approximately 10 to 11 hours for each language pair.

MODEL	DA		TOP-1	
	EN-DE	ZH-EN	EN-DE	ZH-EN
GOOGLETRANSLATE	86.87	79.85	62	114
GPT-4	87.98	79.12	66	108
TOWERINSTRUCT-13B	86.53	69.12	53	56
ALMA-13B-R	84.96	66.02	46	51
TOWERINSTRUCT-7B	83.32	68.66	37	63

Table 1: Human evaluation results: DA scores for all MT systems are high, suggesting that translations are generally of very good quality according to experts.

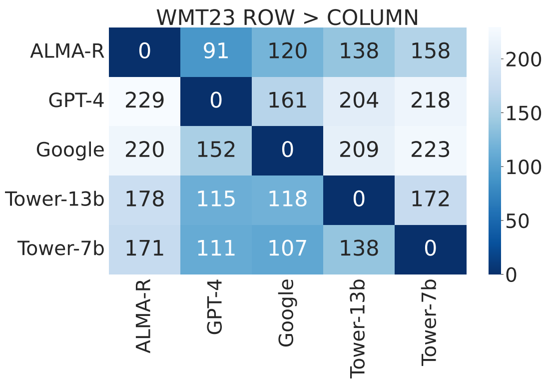

Figure 1: Pairwise Preferences between different Systems: Google and GPT-4 translations are more preferred over open-sourced alternatives.

Overall Quality For EN-DE, DA scores range from 83.32 to 87.98, with no significant difference in the translation quality of different systems, according to the Mann-Whitney test (McKnight and Najab, 2010). On the other hand, DA and Top-1 are significantly better for GPT-4 and GOOGLETRANSLATE models for ZH-EN. Further qualitative analysis shows that for WMT23 ZH-EN, the quality of the source sentences is often poor—up to 25% of source sentences were marked as problematic by the linguist. This suggests there is still room for improvement for open-source models over close-sourced alternatives when generating translations for noisy source texts (Peters and Martins, 2024).

Pairwise Preferences We also report pairwise wins for each model against the other in Fig. 1. GOOGLETRANSLATE and GPT-4 outputs are generally more preferred over open-sourced translation alternatives. Further analysis shows that about 25% and 10% pairs are tied for equal preferences for ZH-EN and EN-DE respectively, further validating close translation quality amongst alternatives. Taken together, these results show that all the evaluated MT systems generate high-quality translations.

METRIC	EN-DE				ZH-EN			
	P	S	TAU	PRECISION@1	P	S	TAU	PRECISION@1
COMETKIWI-XL	0.275	0.272	0.229	**47.0**	0.332	0.336	0.289	42.9
COMETKIWI-XXL	0.253	0.238	0.198	43.9	0.342	0.346	0.279	46.6
XCOMET-XL	0.334	0.300	0.249	41.9	**0.456**	0.410	**0.342**	44.5
XCOMET-XXL	0.316	0.312	0.252	44.4	0.343	0.410	0.340	44.0
METRICX-23-L	0.238	0.238	0.191	37.9	0.428	0.409	0.328	42.4
METRICX-23-XL	0.270	0.245	0.206	39.4	0.417	0.410	**0.342**	45.5
XCOMET-XL+XXL	**0.341**	**0.329**	**0.270**	**47.0**	0.434	**0.411**	0.336	**48.7**
COMETKIWI-XL+XXL	0.273	0.252	0.211	43.9	0.347	0.357	0.290	41.4
XCOMET+KIWI-XXL	0.286	0.271	0.223	45.5	0.377	0.382	0.304	46.6
COMET-REF	0.331	0.286	0.234	50.5	0.243	0.211	0.169	47.1

Table 2: Correlation and Precision@1 for automatic QE metrics: XCOMET-XL+XXL results in the highest correlation and Precision@1 across the board, outperforming reference-based metric, COMET-REF.

3.3 Evaluating Automatic Metrics

We evaluate the best-performing metrics from the WMT23 QE Shared Task: 1) COMETKIWI (Rei et al., 2023); 2) XCOMET (Guerreiro et al., 2023); 3) METRICX (Juraska et al., 2023) and ensembles of these metrics obtained by averaging the scores from the two metrics: 4) COMETKIWI-XL+XXL 5) XCOMET-XL+XXL and 6) COMETKIWI+XCOMET-XXL.[5]

3.3.1 Metrics for Meta-Evaluation

We report the following scores to assess these metrics in their ability to recover human preferences:

Correlation Following WMT evaluation campaign, we report the Pearson (P), Spearman (S), and Kendall Tau (TAU) correlation of automatic metrics with human judgements over all collected judgments grouped by source.

Precision@1 for the best translation We additionally report the precision of identifying the best hypothesis by an automatic metric as the number of times the metric's ranked best translation is in the set of human-ranked best translations. Note that as we ask linguists to provide the same scores to mark equal preferences over different translations, multiple translations can obtain the highest quality.

3.3.2 Findings

Our main results are summarized in Table 2. The correlation between human judgments and metric scores on these high-quality translations is rather low, suggesting a limited ability to model human preferences between multiple translations for the same source. XCOMET-XL+XXL, an ensemble of XCOMET-XL and XCOMET-XXL, achieves the best Spearman and PRECISION@1 across the board, even outperforming reference-based metric COMET (Rei et al., 2020) on this task. Hence, we use this metric to induce preference judgments in our dataset in §4. Designing metrics that accurately reflect these quality preferences remains an open challenge. The dataset collected in our study can potentially be used to benchmark new metrics, which we leave for future work.

[5]We refer the reader to the original papers for each metric for more details about the training and architecture.

4 MT-PREF Dataset

Building on the findings from §3, we create our preference dataset using XCOMET-XL+XXL. We discuss the choice of the text and models in §4.1, followed by the method for inducing and selecting preference pairs from the dataset in §4.2.

4.1 Data and MT Systems

We collect source segments from REDPAJAMA (Computer, 2023) for English, German, French, Spanish, and Italian, and use MC4 (Raffel et al., 2019) for the remaining languages: Portuguese, Russian, Chinese, French, and Korean. Approximately 1000 segments published after July 2022 were extracted and filtered for each language using the perplexity score available in the original REDPAJAMA and MC4 collections. The perplexity thresholds vary across languages and were defined after manual checks on the filtered segments, avoiding non-fluent segments with repetitive patterns such as sequences of numbers, non-alphanumeric characters, and repeated words, among others.

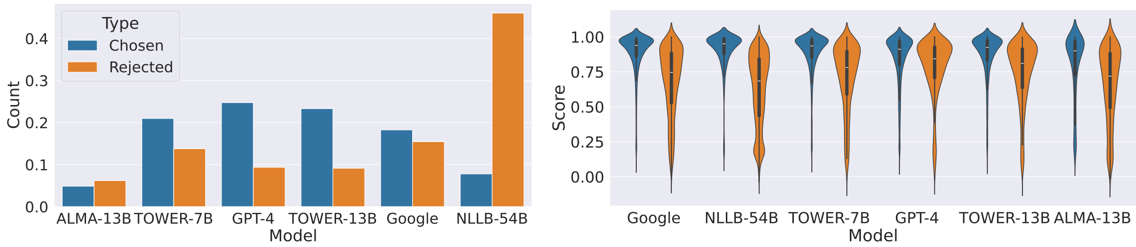

Figure 2: Distribution of counts and scores for the chosen (y_+) and rejected (y_-) hypotheses across models.

We generate translation outputs using greedy decoding from six diverse models varying in architecture (encoder-decoder and decoder-only), model sizes (7B, 13B, 54B), and output quality (see Figure 2). Specifically, we use 1) NLLB-54B (Costa-jussà et al., 2022), 2) ALMA-13B (Xu et al., 2023), 3) GPT-4, 4) GOOGLETRANSLATE, and 5) TOWER models (TOWERINSTRUCT-13B and TOWERINSTRUCT-7B). A detailed description of MT systems is provided in the Appendix B. We generate translations using all models for all directions EN $\Leftrightarrow$ {DE, FR, PT, NL, KO, ZH, RU, ES, IT}, with two exceptions. For ALMA-13B, we only generate outputs for supported language pairs (EN $\Leftrightarrow$ {DE, ZH, RU}) and discard translations for {ZH, KO} $\rightarrow$ EN from NLLB-54B due to inferior quality and frequent hallucinations.

4.2 Creating Preferences

For each source sentence x, we have up to six translation options $\{y_j\}_{j=1}^6$. Our goal is to get preference triples of source (x), a preferred/chosen hypothesis (y_+), and a less preferred/rejected hypothesis (y_-). We use an automatic quality estimation metric $\mathcal{M}$ to create this dataset of preference triples $\mathcal{D} = \{(x, y_+, y_-)\}$ and resort to a simple criterion that obtains the maximum discrepancy under $\mathcal{M}$. We first measure the translation quality scores for each pair (x, y_j), resulting in the scores $s = \{s_j\}_{j=1}^6$. We then select the best and the worst translation hypotheses from the ranked list induced by the scores, s, *i.e.* $y_+ = y_{\arg\max_j(s_j)}$ and $y_- = y_{\arg\min_j(s_j)}$. This results in a unique preference triplet for each source sentence.

5 Experimental Settings

We use the MT-PREF dataset to align MT models with translation preferences (§4) and compare several preference learning methods detailed in §2.

Training Data The MT-PREF dataset contains 18k instances with approximately 1k examples for each translation direction. The counts of the chosen and the rejected hypotheses from each model and the distribution of metric scores are shown in Fig. 2. The NLLB-54B model accounts for most of the rejected hypotheses (∼46%), whereas the chosen hypotheses are more equally distributed across the GPT-4, GOOGLETRANSLATE, and the TOWER models, illustrating consistent and higher-quality translations generated by these models.

Evaluation We evaluate finetuned models on the WMT23 test set (EN $\leftrightarrow$ {DE, RU, ZH}) and the FLORES dev-test set (EN $\leftrightarrow$ DE, RU, ZH, ES, FR, PT, NL, IT, KO) using TOWER-EVAL.[6] We report system-level translation quality using CHRF (Popović, 2015), COMET, and XCOMET-XL. We cluster system performance using the Wilcoxon rank-sum test ($p < 0.05$) with COMET as the primary metric. Rank ranges, denoted by $[l+1, n-w]$, indicate the number of systems a particular system underperforms or outperforms, where l represents the number of losses, n is the total number of systems, and w is the number of systems that the system in question significantly outperforms (Kocmi et al., 2023). We compare the models' accuracy (% ACC.) for selecting the best-over-worst hypothesis with the model's likelihood on the human preferences (§3) after finetuning on MT-PREF.

Model Configurations We finetune TOWERINSTRUCT-7B using preference optimization methods detailed in §2 with the following configurations:

- SFT: a baseline model supervised finetuned on the chosen or the most preferred response.
- DPO_{sft}: model trained with π_{ref}=SFT in Eq. 2.
- DPO_{base}: base model directly finetuned with DPO, *i.e.* π_{ref}=TOWERINSTRUCT-7B.
- DPO_{base}+SFT: base model finetuned with a combination of DPO and SFT regularization, *i.e.* $\mathcal{L}_{\text{DPO}}(x, y_\pm; \pi_\theta, \pi_{\text{ref}}) + \lambda\mathcal{L}_{\text{NLL}}(x, y_+; \theta)$.

[6] https://github.com/deep-spin/tower-eval

MODEL	EN-XX				XX-EN				% ACC.
	CHRF	COMET	XCOMET-XL	RANK	CHRF	COMET	XCOMET-XL	RANK	
TOWERINSTRUCT-7B	52.25	84.32	85.32	9-12	58.87	82.77	88.77	9-12	53.25
+ SFT	**53.29**	84.26	85.11	9-12	59.30	82.79	89.16	5-12	58.50
+ DPO_{sft}	53.27	84.85	85.63	5-9	**59.86**	**83.18**	89.56	3-11	59.25
+ DPO_{base}	49.90	84.64	86.14	9-12	58.34	83.05	**89.73**	4-12	59.50
+ DPO_{base}+SFT	52.42	84.99	86.37	5-8	59.43	83.16	89.60	3-11	58.25
+ CPO	52.95	**85.05**	**86.43**	5-8	59.62	83.14	89.70	3-11	**59.50**
TOWERINSTRUCT-13B	54.15	85.17	86.55	5-8	59.86	83.18	89.33	4-12	59.50
+ CPO	54.45	85.59	87.22	3-4	60.55	83.49	89.98	2-7	60.25
ALMA-13B-R	47.57	84.95	87.27	8-12	58.79	83.12	89.43	5-12	50.00
GPT-3.5	56.38	85.56	86.92	3-4	60.92	83.48	90.00	2-9	-
GPT-4	56.94	86.01	87.43	2	61.33	83.69	**90.34**	2-4	-
GOOGLETRANSLATE	**60.43**	**86.44**	**87.53**	1	**62.05**	**84.07**	89.83	1	-

Table 3: Comparing PO methods on WMT23: Both CPO and DPO_{base}+SFT result in significant improvement in translation quality, closing the gap with TOWERINSTRUCT-13B.

- CPO: model finetuned with the objective in Eq. 4.

We also compare the aligned models against TOWERINSTRUCT-13B, GPT-4, ALMA-13B-R and GOOGLETRANSLATE models. All training details are provided in Appendix D.

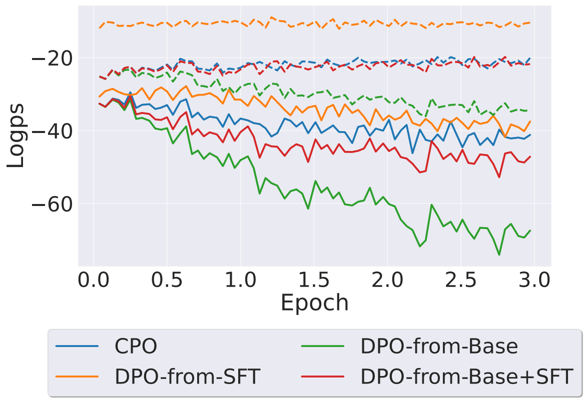

Figure 3: Log probabilities for chosen (- - -) and rejected (—) hypotheses during training across PO methods: DPO_{base} reduces the likelihood for both chosen and rejected responses, resulting in reduced output quality.

6 Results

We first present the results of comparing several PO methods (§2) in Table 3 on the WMT23 and FLORES datasets. Scores are averaged for from-English (EN-XX) and to-English (XX-EN) translation directions. Results for individual language pairs are shown in Appendix E. We then compare preference learning on MT-PREF against an existing preference dataset (§6.2), followed by an ablation on the impact of the dataset size on the final translation quality (§6.3).

6.1 Comparing PO Algorithms

SFT results in limited translation quality gains. SFT on the *chosen* response from the MT-PREF dataset improves CHRF over TOWERINSTRUCT-7B on EN-XX (+1.04) and XX-EN (+0.43) translation directions, with no significant difference in COMET and XCOMET-XL in EN-XX direction. However, we observe a large gain (+5.25%) in % ACC., suggesting that the model does acquire some ability to distinguish high-quality translations even when trained with best translations only.

Preference learning improves translation quality. Most PO methods improve COMET and XCOMET-XL as well as % ACC. over TOWERINSTRUCT-7B in both directions, showing that aligning LLMs with preferences benefits MT. The translation quality gap between TOWERINSTRUCT-7B and TOWERINSTRUCT-13B by COMET is reduced significantly. Optimizing TOWERINSTRUCT-13B on MT-PREF with CPO further improves translation quality significantly reaching comparable quality to GPT-3.5 and GPT-4 for EN-XX and XX-EN directions respectively. This illustrates that finetuning on MT-PREF can improve translation quality even for larger models.

SFT is necessary to obtain translation quality improvements using DPO. Comparing different variants of DPO (DPO_{sft}, DPO_{base} and DPO_{base}+SFT), we find that either the SFT phase or the SFT regularization is necessary to obtain sig-

Dataset	EN-XX				XX-EN			
	chrF	COMET	xCOMET-XL	Rank	chrF	COMET	xCOMET-XL	Rank
TowerInstruct-7B	56.14	88.51	93.01	4	64.08	88.28	96.20	3-4
+ CPO	**56.70**	**88.81**	**93.71**	2-3	**64.21**	**88.32**	**96.56**	3-4
TowerInstruct-13B	57.17	88.89	93.85	2-3	64.80	88.50	96.44	1-2
+ CPO	**57.79**	**89.15**	**94.30**	1	**64.90**	**88.51**	**96.71**	1-2

Table 4: CPO finetuning using MT-Pref improves translation quality for Tower models on FLORES.

Dataset	Metric	N	EN-XX				XX-EN			
			chrF	COMET	xCOMET-XL	Rank	chrF	COMET	xCOMET-XL	Rank
Tower-7B	-	-	52.25	84.32	85.32	5-7	58.87	82.77	88.77	5-7
MT-Pref	xCOMET-XL+XXL	18k	52.95	**85.05**	**86.43**	1-5	59.62	83.14	**89.70**	1-6
		6k	**52.98**	84.81	85.98	2-6	59.63	83.09	89.46	1-7
	xCOMET+Kiwi–XXL	18k	52.87	84.86	85.90	1-6	**59.86**	**83.15**	89.54	1-6
ALMA-R	xCOMET+Kiwi-XXL	14k	49.87	84.89	86.35	3-7	59.63	83.24	89.47	1-6
		6k	51.02	84.76	85.90	2-7	59.72	**83.15**	89.33	1-6
Tower-13B			54.15	85.17	86.55	1-3	59.86	83.18	89.33	1-7

Table 5: CPO finetuning on ALMA-R-Pref and MT-Pref variants: Preferences induced via xCOMET-XL+XXL on all examples gives the best overall results.

nificant COMET improvements. This also aligns with findings from Tunstall et al. (2023) who show that learning from chat preference datasets fails when skipping the initial SFT stage. Interestingly, DPO_{base} attains the highest % Acc. scores among variants, showing an improved ability to discern but not necessarily generate high-quality translations. We find that as suggested by (Pal et al., 2024), it is indeed because DPO_{base} increases the relative probability between the two classes by decreasing the model's likelihood for both *chosen* and *rejected* translations (see Fig. 3).

Results on FLORES We report the results of aligning TowerInstruct-7B with CPO on Flores in Table 4. On average, the translation quality of the base models, TowerInstruct-7B and TowerInstruct-13B, improves with alignment tuning across the board according to all metrics, with TowerInstruct-7B reaching close COMET and xCOMET-XL scores to TowerInstruct-13B, despite being 2 times smaller in size.

In gist, we show that CPO results in the best-aligned TowerInstruct-7B, matching translation quality with TowerInstruct-13B on both WMT23 and FLORES benchmarks. We next compare preference optimization using CPO on MT-Pref against existing preference datasets.

6.2 MT-Pref Vs. ALMA-R-Pref

Xu et al. (2024) use the FLORES-200 development and test datasets to create a preference dataset, ALMA-R-Pref. For each source sentence in the corpus, they take the human-written reference, and outputs from ALMA-13B-R and GPT-4 models, and induce preferences using an ensemble of xCOMET-XXL and CometKiwi-XXL metrics. We note that this metric ensemble attains similar or lower correlation scores compared to the best individual metrics on both language pairs as shown in Table 2. We compare the translation quality of the resulting models when aligned with MT-Pref and ALMA-R-Pref preference datasets in Table 5.[7]

Training on ALMA-R-Pref preference dataset improves neural metrics but significantly hurts chrF compared to the base model, TowerInstruct-7B.[8] Our analysis shows that finetuning on the ALMA-R-Pref dataset increases the output length significantly. This could be due to the inherent bias in the dataset where the *chosen* responses, typically by GPT-4 (45%), are on average longer than the *rejected* responses.[9] This has im-

[7] We do not compare with MAPLE (Zhu et al., 2024) due to lack of open access to this dataset.

[8] A difference of 2.4 chrF points is considered significant with 87% accuracy (Kocmi et al., 2024).

[9] The difference in the length of *chosen* and *rejected* translations in the training dataset is also significant according to an independent t-test with a p-value of 0.01.

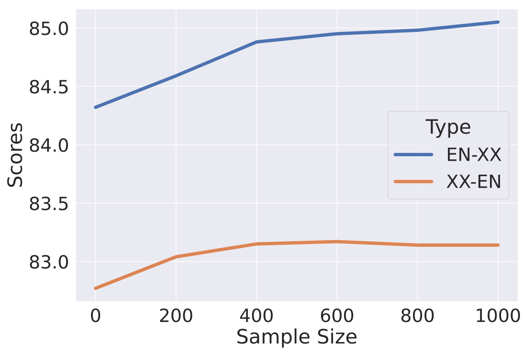

Figure 4: COMET with varying size of the preference dataset: EN-XX continues to benefit from more samples.

portant implications for the creation and modeling of preferences – when a model is too frequently "preferred" in a dataset, it can lead to the distillation of that model's characteristics and it is unclear to what extent humans prefer these distilled features.

TOWERINSTRUCT-7B finetuned on equal-sized ALMA-R-PREF and MT-PREF datasets score close on neural metrics, with a difference of 1.96 points on CHRF. As our preference dataset considers outputs from multiple models with diverse styles, we do not distill any such model-specific biases. Furthermore, aligning on preferences induced via XCOMET-XL+XXL yields slightly better COMET score on EN-XX direction over preferences with XCOMET+KIWI-XXL, further validating the importance of inducing preferences using metrics guided by human knowledge.

6.3 Impact of the Size of Preference Datasets

One advantage of our approach is that we can scale the size of preference datasets as necessary as preferences are induced using an automatic QE metric. To understand whether this is indeed beneficial, we conduct an ablation where we vary the number of unique source samples per language pairs as: {200, 400, 600, 800, 1000} and align TOWERINSTRUCT-7B on the resulting preference dataset using CPO. Fig. 4 shows the results: while the improvement in quality for XX-EN plateaus with just 400 samples per language direction, COMET continues to improve for EN-XX suggesting that adding more data might benefit translations from English to other language pairs. This aligns with the fact that the model is exposed to relatively fewer non-English texts during pretraining and hence benefits more from any additional dataset on these languages.

7 Related Work

LLMs for MT Earlier works exploring LLMs to perform MT study prompting techniques to generate translations (Hendy et al., 2023; Zhang et al., 2023a; Vilar et al., 2023) with research focusing on selecting high-quality and relevant examples as demonstrations to incorporating external knowledge mimicking human-like translation strategies (He et al., 2023). More recently, several works have proposed finetuning LLMs to improve the translation quality (Zhang et al., 2023b; Alves et al., 2023), resulting in specialized models that attain competitive performance to state-of-the-art production level translation systems (Xu et al., 2023; Alves et al., 2024). Across all methods, the quality of the data used for training is paramount to the finetuning methods. Therefore, in this work, we focus on curating a high-quality translation preference dataset using metrics that closely reflect true human translation preferences and outputs generated from a diverse set of high-quality MT systems.

Quality Feedback for MT Using feedback from automatic metrics for MT or human quality assessment has been an active area of research through the past decade. This quality signal is either utilized during training (Shen et al., 2016; Wieting et al., 2019; Yang et al., 2023; He et al., 2024; Gulcehre et al., 2023; Nguyen et al., 2017; Kreutzer et al., 2018, 2020) or decoding (Freitag et al., 2022; Fernandes et al., 2022; Farinhas et al., 2023) or for modeling translation preferences in the dataset directly (Xu et al., 2024; Zhu et al., 2024). Similar to Xu et al. (2024), we use automatic metrics to induce preferences in the dataset but with the additional validation that the chosen metric indeed reflects human quality expectations and with translations generated from diverse MT systems.

8 Conclusion

We present MT-PREF, a high-quality translation preference dataset, curated by combining the strengths of human evaluation and automatic metrics. The dataset includes metric-induced preferences from strong MT models across 18 language directions with new source sentences mined post-2022. Aligning state-of-the-art decoder-only LLMs on this preference dataset using existing aligning tuning algorithms improves translation quality. Furthermore, the aligned models are also better at modeling human preferences of translation quality.

Limitations

We note a few limitations of our work. We evaluate the translation quality of the finetuned models primarily using automatic metrics. While we validate that they can indeed provide a reasonable signal to differentiate quality at the system level (See Appendix C), it requires a human evaluation to confirm whether and to what extent the aligned models match human preferences. Furthermore, we use existing QE metrics that can be sensitive to the domain of the datasets (Zouhar et al., 2024). However, as the QE metrics continue to improve, our approach allows to substitute the preferences with that induced by a better QE metric. Finally, we do not handle tied preferences in translation quality and always induce a strict preference order. Incorporating neutral preferences between translations can help the model focus on attributes that truly improve quality over stylistic preferences; we leave the investigation of this phenomenon to future work. We note that our dataset can be used to design better QE metrics for ranking translations, inducing preferences using new criteria, and employing better optimization methods.

Potential Risks

Large language models may carry the potential risk of generating fluent and hallucinated content. When the users do not know the target or the source language, they might trust the generated translation without further verification (Martindale and Carpuat, 2018). And while our approach is driven toward making the model aware of translations of varying quality during finetuning, the coverage is limited to the supported language pairs. Users should exercise caution and seek verification from additional sources where possible when using LLMs on real-world applications.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Duarte Alves, Nuno Guerreiro, João Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and Andre Martins. 2023. Steering large language models for machine translation with finetuning and in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148, Singapore. Association for Computational Linguistics.

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9432–9452, Toronto, Canada. Association for Computational Linguistics.

Together Computer. 2023. Redpajama: an open dataset for training large language models.

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

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.

António Farinhas, José de Souza, and Andre Martins. 2023. An empirical study of translation hypothesis ensembling with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11956–11970, Singapore. Association for Computational Linguistics.

Christian Federmann. 2018. Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Yang Gao, Dana Alon, and Donald Metzler. 2024. Impact of preference noise on the alignment performance of generative language models. *arXiv preprint arXiv:2404.09824*.

Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection. *arXiv preprint arXiv:2310.10482*.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, et al. 2023. Reinforced self-training (rest) for language modeling. *arXiv preprint arXiv:2308.08998*.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. Exploring human-like translation strategy with large language models. *arXiv preprint arXiv:2305.04118*.

Zhiwei He, Xing Wang, Wenxiang Jiao, Zhuosheng Zhang, Rui Wang, Shuming Shi, and Zhaopeng Tu. 2024. Improving machine translation with human feedback: An exploration of quality estimation as a reward model. *arXiv preprint arXiv:2401.12873*.

Joey Hejna, Rafael Rafailov, Harshit Sikchi, Chelsea Finn, Scott Niekum, W Bradley Knox, and Dorsa Sadigh. 2023. Contrastive prefence learning: Learning from human feedback without rl. *arXiv preprint arXiv:2310.13639*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. 2012. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8):2.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. Navigating the metrics maze: Reconciling score magnitudes and accuracies. *arXiv preprint arXiv:2401.06760*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Julia Kreutzer, Nathaniel Berger, and Stefan Riezler. 2020. Correct me if you can: Learning from error corrections and markings. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 135–144, Lisboa, Portugal. European Association for Machine Translation.

Julia Kreutzer, Shahram Khadivi, Evgeny Matusov, and Stefan Riezler. 2018. Can neural machine translation be improved with user feedback? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 92–105, New Orleans - Louisiana. Association for Computational Linguistics.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2024. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. *Transactions of the Association for Computational Linguistics*, 12:576–592.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2023. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. *arXiv preprint arXiv:2312.15685*.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

Marianna Martindale and Marine Carpuat. 2018. Fluency over adequacy: A pilot study in measuring user trust in imperfect mt. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 13–25.

Stephen Mayhew, Klinton Bicknell, Chris Brust, Bill McDowell, Will Monroe, and Burr Settles. 2020. Simultaneous translation and paraphrase for language education. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 232–243, Online. Association for Computational Linguistics.

Patrick E McKnight and Julius Najab. 2010. Mann-whitney u test. *The Corsini encyclopedia of psychology*, pages 1–1.

Tetsuro Morimura, Mitsuki Sakamoto, Yuu Jinnai, Kenshi Abe, and Kaito Air. 2024. Filtered direct preference optimization. *arXiv preprint arXiv:2404.13846*.

Khanh Nguyen, Hal Daumé III, and Jordan Boyd-Graber. 2017. Reinforcement learning for bandit neural machine translation with simulated human feedback. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1464–1474, Copenhagen, Denmark. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. 2024. Smaug: Fixing failure modes of preference optimisation with dpo-positive. *arXiv preprint arXiv:2402.13228*.

Ben Peters and André FT Martins. 2024. Did translation models get more robust without anyone even noticing? *arXiv preprint arXiv:2403.03923*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1692, Berlin, Germany. Association for Computational Linguistics.

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

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU:training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. Association for Computational Linguistics.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.

Guangyu Yang, Jinghong Chen, Weizhe Lin, and Bill Byrne. 2023. Direct preference optimization for neural machine translation with minimum bayes risk decoding. *arXiv preprint arXiv:2311.08380*.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2023. Tim: Teaching large language models to translate with comparison. *arXiv preprint arXiv:2307.04408*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, pages 41092–41110. PMLR.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023b. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481, Singapore. Association for Computational Linguistics.

Dawei Zhu, Sony Trenous, Xiaoyu Shen, Dietrich Klakow, Bill Byrne, and Eva Hasler. 2024. A preference-driven paradigm for enhanced translation with large language models.

Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024. Fine-tuned machine translation metrics struggle in unseen domains. *arXiv preprint arXiv:2306.07899*.

A Annotation Guidelines and Interface

Task Overview This task involves evaluating five translations of a source text and assigning a quality rating to each translation based on its overall quality and adherence to the source content. You will need to consider the accuracy, fluency, and overall quality when assessing the different translations.

Annotation Scale Each translation is evaluated on a continuous scale of 0-6 with the quality levels described as follows:

- 6: Perfect Meaning and Grammar: The meaning of the translation is completely consistent with the source and the surrounding context (if applicable). The grammar is also correct.
- 4: Most Meaning Preserved and Few Grammar Mistakes: The translation retains most of the meaning of the source. It may have some grammar mistakes or minor contextual inconsistencies.
- 2: Some Meaning Preserved: The translation preserves some of the meaning of the source but misses significant parts. The narrative is hard to follow due to fundamental errors. Grammar may be poor.
- 0: Nonsense/No meaning preserved: Nearly all information is lost between the translation and source. Grammar is irrelevant.

You can scroll up or down to see all the other translation outputs from the different systems. Figure 5 shows the interface when comparing and evaluating five translations. While each translation is evaluated independently, these translations can also be ranked based on the difference in their absolute scores. It is perfectly valid to give the same score to multiple translations if you believe they are of the same overall quality.

Other Details We hired native speakers of Chinese and German for this task (both females) and they were compensated at $20 per hour.

B MT Systems

We use the following MT systems:

1. **NLLB-54B** (Costa-jussà et al., 2022) is a 54B encoder-decoder multilingual translation model, based on a sparsely gated Mixture of Experts (MoE) approach. It covers 202 languages, supporting translation for many low-resource languages.
2. **TOWERINSTRUCT-13B and TOWERINSTRUCT-7B** are 13B and 7B decoder-only LLMs, trained to optimize quality on multiple tasks present in translation workflows. The model is continued pretrained from LLAMA 2 (Touvron et al., 2023) checkpoints on a multilingual mixture of monolingual and parallel data, followed by finetuning on instructions relevant to translation processes.
3. **ALMA-13B** (Xu et al., 2023) is a 13B decoder-only model specialized for MT via continued pretaining, followed by instruction tuning on a small but high-quality parallel dataset. Unlike TOWERINSTRUCT models, the continued pretraining phase only explores monolingual data, and the instruction tuning is performed with an MT dataset only.
4. **ALMA-13B-R** (Xu et al., 2024) is a 13B decoder-only model obtained by finetuning ALMA-13B with ALMA-R-PREF using CPO.
5. **GPT-4** (Achiam et al., 2023) is prompted in a zero-shot fashion, following Hendy et al. (2023), to generate translations using the prompt:
   Translate this sentence from [source language] to [target language]:
   Source: [source sentence]
   Target:
6. **GOOGLETRANSLATE** is the basic version of the Translate API v2 accessed on 2024-03-04.[10]

[10] https://translation.googleapis.com/language/translate/v2

MODEL	EN-DE				ZH-EN			
	CHRF	COMET	XCOMET-XL	DA	CHRF	COMET	XCOMET-XL	DA
GOOGLETRANSLATE	68.83 (1)	0.854 (1)	0.941 (1)	86.87 (2)	49.40 (1)	0.810 (1)	0.884 (1)	79.85 (1)
GPT-4	68.50 (2)	0.848 (2)	0.932 (3)	87.98 (1)	45.95 (2)	0.799 (2)	0.877 (2)	79.12 (2)
TOWERINSTRUCT-13B	66.45 (3)	0.843 (3)	0.931 (4)	86.53 (3)	45.29 (3)	0.794 (3)	0.866 (3)	69.12 (3)
ALMA-13B-R	59.92 (5)	0.836 (4)	0.935 (2)	84.96 (4)	44.72 (4)	0.793 (4)	0.858 (5)	66.02 (5)
TOWERINSTRUCT-7B	64.61 (4)	0.830 (5)	0.918 (5)	83.32 (5)	43.77 (5)	0.790 (5)	0.860 (4)	68.66 (4)
PAIRWISE-ACC	8/10	9/10	7/10	-	9/10	9/10	10/10	-

Table 6: Automatic Evaluation - System Level for reference-based metrics. Ranks represent the ordering based on averaged DA scores.

C System-level Correlation

Table 6 shows the system-level translation quality scores assigned by reference-based metrics: CHRF, COMET, and XCOMET-XL for all five models and their induced system-level rankings. For both directions, COMET results in 90% agreement with human judgments, confirming its accuracy in rating high-quality systems and hence we use COMET as the primary metric for ranking different systems.

D Training Details

Hyperparameters We finetune TOWERINSTRUCT-7B and TOWERINSTRUCT-13B models (Alves et al., 2024) using the TRL library (von Werra et al., 2020) with a batch size of 64, a maximum output length of 256, a learning rate of 5×10^{-7} and a warm-up ratio of 0.1. The model is finetuned using different preference algorithms (§2) for 3 epochs with RMSProp optimizer (Hinton et al., 2012). For SFT, following (Tunstall et al., 2023), we finetune the base model for one epoch with a learning rate of 1×10^{-5} using Adam optimizer (Kingma and Ba, 2014). We use greedy decoding to generate translation hypotheses using the aligned models. All our models are trained on two Nvidia A100 GPUs. Training takes approximately four to five hours to converge.

E Results by WMT23 Language Direction

We report results comparing preference optimization methods when trained with MT-PREF on individual language pairs using COMET, CHRF and XCOMET-XL in Tables 7, 8 and 9 respectively.

MODEL	EN-DE	EN-ZH	EN-RU	DE-EN	ZH-EN	RU-EN
TOWERINSTRUCT-7B	83.25	84.98	84.72	85.25	80.15	82.90
+ SFT	83.01	85.47	84.29	85.25	80.25	82.86
+ DPO_{sft}	83.83	85.81	84.91	85.66	80.72	83.17
+ DPO_{base}	83.73	84.64	85.55	85.25	80.60	83.30
+ DPO_{base}+SFT	83.86	85.65	85.46	85.53	80.69	83.26
+ CPO	83.92	85.74	85.49	85.47	80.79	83.17
TOWERINSTRUCT-13B	84.02	85.97	85.52	85.60	80.71	83.23
+ CPO	84.53	86.32	85.91	85.72	81.25	83.49
ALMA-13B-R	84.03	84.97	85.85	85.54	80.55	83.28
GPT-3.5	84.61	86.70	85.38	85.91	81.52	83.02
GPT-4	84.89	87.08	86.07	86.17	81.27	83.63
GOOGLETRANSLATE	84.77	88.09	86.45	86.24	82.19	83.78

Table 7: COMET on WMT23 dataset comparing PO methods when trained with MT-PREF.

MODEL	EN-DE	EN-ZH	EN-RU	DE-EN	ZH-EN	RU-EN
TOWERINSTRUCT-7B	65.74	37.34	53.66	67.80	49.91	58.89
+ SFT	65.76	40.16	53.95	67.93	50.89	59.08
+ DPO_{sft}	66.16	39.56	54.10	68.57	51.61	59.38
+ DPO_{base}	64.23	33.02	52.43	66.61	50.25	58.15
+ DPO_{base}+SFT	65.90	37.59	53.78	67.97	50.89	59.42
+ CPO	66.22	38.68	53.96	68.25	51.31	59.30
TOWERINSTRUCT-13B	66.90	40.62	54.95	68.47	51.22	59.89
+ CPO	67.36	40.74	55.24	69.03	52.35	60.28
ALMA-13B-R	60.38	32.14	50.19	66.30	51.28	58.79
GPT-3.5	68.38	45.25	55.50	69.21	53.78	59.77
GPT-4	69.30	45.67	55.86	69.91	53.37	60.70
GOOGLETRANSLATE	69.08	52.99	59.21	70.28	55.15	60.72

Table 8: CHRF on WMT23 dataset comparing PO methods when trained with MT-PREF.

MODEL	EN-DE	EN-ZH	EN-RU	DE-EN	ZH-EN	RU-EN
TOWERINSTRUCT-7B	84.44	83.77	87.75	89.07	85.02	92.23
+ SFT	84.48	83.67	87.19	89.24	85.75	92.48
+ DPO_{sft}	84.98	84.13	87.78	89.62	86.22	92.84
+ DPO_{base}	85.17	83.78	89.47	89.54	86.41	93.23
+ DPO_{base}+SFT	85.24	84.67	89.20	89.51	86.33	92.95
+ CPO	85.33	84.98	88.97	89.51	86.70	92.88
TOWERINSTRUCT-13B	85.42	85.17	89.05	89.41	85.81	92.77
+ CPO	86.13	85.80	89.74	89.86	86.86	93.21
ALMA-13B-R	86.09	84.81	90.91	89.24	86.14	92.92
GPT-3.5	86.62	85.16	88.99	89.80	87.23	92.98
GPT-4	86.72	85.59	89.98	89.92	87.43	93.68
GOOGLETRANSLATE	85.76	86.73	90.11	89.37	86.93	93.20

Table 9: XCOMET-XL on WMT23 dataset comparing PO methods when trained with MT-PREF.

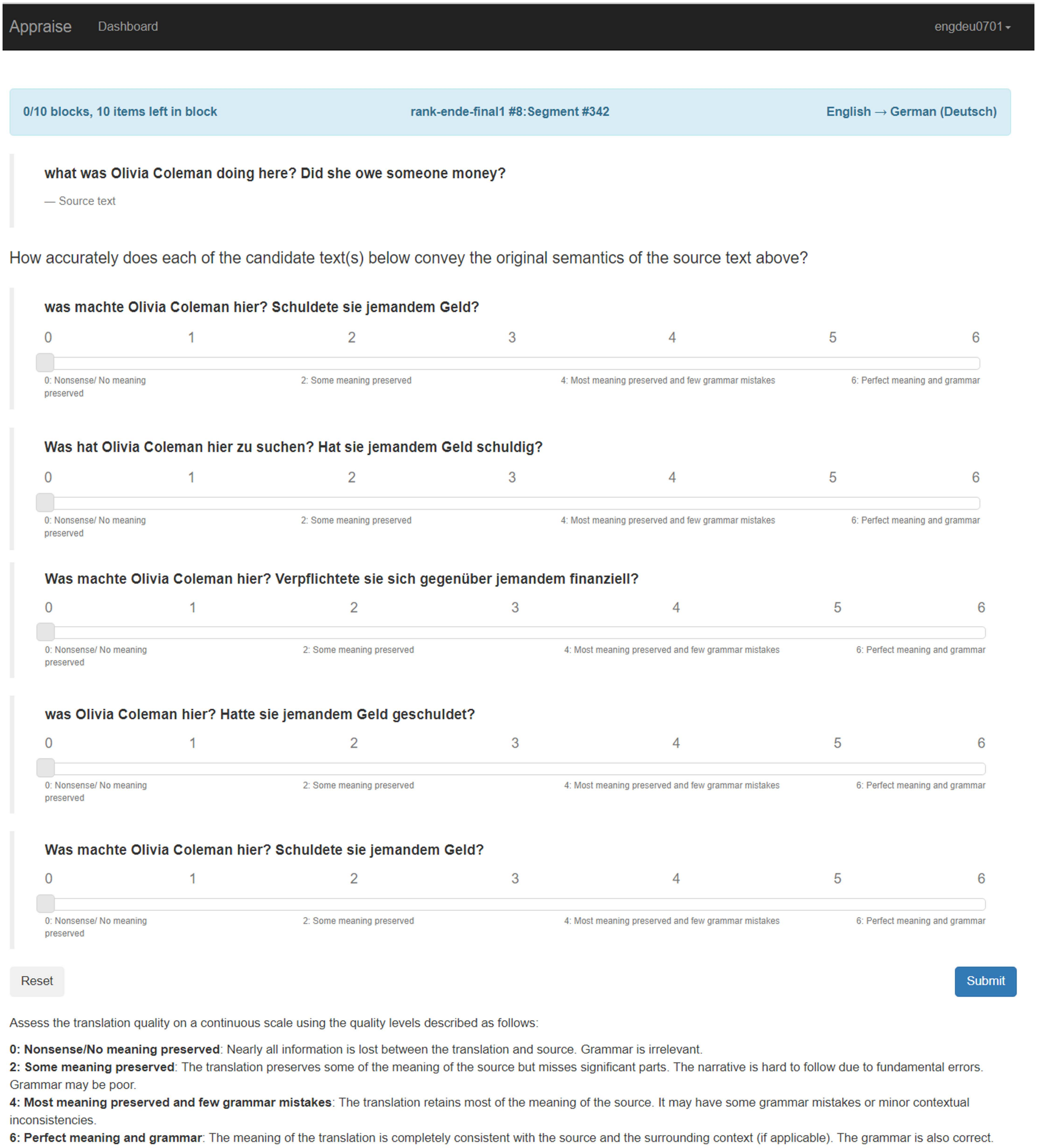

Figure 5: Annotation Interface.