Cross-lingual Human-Preference Alignment for Neural Machine Translation with Direct Quality Optimization

Kaden Uhlig[1], **Joern Wübker**[1], **John DeNero**[1], **Raphael Reinauer**[2]*
[1]LILT, [2]Amazon
{kaden.uhlig,joern,john}@lilt.com, raphada@amazon.de

Abstract

Reinforcement Learning from Human Feedback (RLHF) and derivative techniques like Direct Preference Optimization (DPO) are task-alignment algorithms used to repurpose general, foundational models for specific tasks. We show that applying task-alignment to neural machine translation (NMT) addresses an existing task–data mismatch in NMT, leading to improvements across all languages of a multilingual model, even when task-alignment is only applied to a subset of those languages. We do so by introducing Direct Quality Optimization (DQO), a variant of DPO leveraging a pre-trained translation quality estimation model as a proxy for human preferences, and verify the improvements with both automatic metrics and through human evaluation.

1 Introduction

For many natural language generation (NLG) tasks, aligning models to human preferences has led to large performance gains (Ziegler et al., 2020). A strong motivation for this alignment step is that much of the data on which the model was originally trained – internet text – is useful for language generation in general but does not match the desired output for the task. Neural machine translation (NMT) models have not involved alignment to human preferences, in part because of the assumption that supervised training data for NMT does match the desired output of the translation task. However, we show the existence of a mismatch between the NMT task and typical training data.

Machine translation is unusual among NLG tasks in that task-relevant supervised training data – text paired with its translation – is plentiful and publicly available. One might expect that with such a large amount of task-relevant training data, there would be no need for task-alignment. However, we identify an exhaustive list of reasons why training examples in a parallel corpus diverge from the desired output in meaningful ways (see Section 2.2).

Machine translation is also unusual in that human preference data has been collected and published for a large number of systems, and translation quality estimation (QE) is an active research area that has benefited greatly from recent advances in large language models. We introduce a method for using quality estimation models, which themselves are trained from human preference data, in order to perform NMT task alignment. Our method, Direct Quality Optimization (DQO), is a batched online variant of Direct Preference Optimization (DPO) (Rafailov et al., 2023) that uses a QE model as a proxy for human preference.

We show that DQO improves translation quality in terms of BLEU, COMET22, CometKiwi22, and BLEURT, and leads to a reduction in translation errors in a human evaluation using the Multidimensional Quality Metric framework (MQM) (Lommel et al., 2014; Freitag et al., 2021).

We make three notable observations when applying DQO to a multilingual model:

1. Task alignment increases task performance and human preference while also increasing the distance between the model's output distribution and the training data distribution.
2. Improvements carry over to held-out languages and language families, which were not contained in the data used for DQO.
3. Improvements in held-out languages are not limited to general behaviors required by the translation task (e.g. avoiding omissions), but include improving language-specific linguistic features not seen in the DQO alignment data, such as correctly transliterating named entities in Latvian.

While we attribute much of the performance in held-out languages to transfer learning of general behaviors required by the translation task

*Contributions made prior to joining Amazon.

Proceedings of the Tenth Workshop on Machine Translation (WMT), pages 31–51
November 8–9, 2025.

(e.g. avoiding omissions), the language-specific improvements in held-out languages cannot be explained by transfer learning.

Instead, these results suggest that DQO does not only increase the likelihood of the features present in its task alignment data, but also focuses the model on human preference features that it already learned during supervised training.

2 The Task–Data Mismatch in NMT

2.1 Task: Human-Preferred Translations

Like many NLG tasks, NMT is an open-ended problem, with multiple valid outputs for any given input, each preferred more or less by humans depending on a variety of factors, including adequacy, fluency, context, tone, style, and many other subtle features.

Because of this, the task of NMT cannot be reduced to producing valid translations, nor human-like translations, but instead requires generating human-preferred translations – those judged as at least as good as all other valid translations.

2.2 Training Data Mismatch

The supervised training data used in NMT comes from a variety of sources, each with notable differences from the task distribution of human-preferred translations.

Web Data Mining. A large portion of parallel data is mined from massive collections of web documents, using automated methods to align source and target language segments – e.g. the ParaCrawl (Bañón et al., 2020) and CCMatrix (Schwenk et al., 2021b) datasets. This process may capture human translations, text written independently in both the source and target languages on the same topic, or the output of other MT models. One prominent cause of task–data mismatch in automatically aligned sentence pairs is semantic misalignment. Kreutzer et al. (2022) found semantic misalignment in 15% (ParaCrawl) and 32% (CCMatrix) of sentence pairs in a manual quality audit.

The simplest form is complete semantic misalignment, when the source and target segments are completely unrelated. This certainly contributes to any task–data mismatch, but such pairs are easy to detect with tools such as BiCleaner (Ramírez-Sánchez et al., 2020) or reference-free quality evaluation models such as CometKiwi22 (Peter et al., 2023).

Unfortunately, slight semantic misalignments of source and target are both more prevalent and much more difficult for state-of-the-art filtering systems to detect (Meng et al., 2024). These may include subtle yet significant differences in meaning, factual differences in numbers or names, additions and omissions, and the accompanying losses in translation adequacy. In addition, these segments often still contain useful information that may help the model learn (Meng et al., 2024).

Accidental Inclusion of Machine Translated Content. Web data may also include the outputs of other machine translation models, including neural, statistical and dictionary-based methods of varying quality. The impact of training on low quality machine translations is clear, however even good NMT systems' outputs differ significantly enough from natural text that classifiers can detect machine translated text with high accuracy – and even predict which machine translation system was used to translate a given text (La Morgia et al., 2023).

Recent research suggests that up to 57% of translations mined from the web are multi-way parallel, meaning parallel translations of a segment can be found in more than two languages, and demonstrates a strong correlation between multi-way parallelism and low quality translations likely to be machine translated (Thompson et al., 2024). The authors also found that multi-way parallel translations follow a distinct distribution, focused on low-quality content typically used for search engine optimization.

Translator Skill Level. Another source of task–data mismatch in human translations is the fact that translators differ in skill level (Albir, 2017). This implies that not all human translations will be equally preferred by humans.

Achieving mean human quality in translations is not the task of NMT as defined in Section 2.1. We propose that neither is *maximum* human quality. In theory it is conceivable that humans prefer machine-generated translations over even the best human-generated translations. Therefore, we do not want finite human skill to impose an upper limit on translation quality.

Translationese. Another common issue is a phenomenon known as *translationese*, the observation that human-translated text in a given language differ in distribution from text written independently in that language. Specifically, translated text shows signs of interference from the source language's grammar, word order and word choice, as well as source-language-independent effects of the translation process itself, such as simplification

Model	Lang.	FLORES+ devtest				NTREX			
		BLEURT	COMET22	CometKiwi22	BLEU	BLEURT	COMET22	CometKiwi22	BLEU
Baseline	All	0.7614	0.8741	0.8387	34.19	0.7016	0.8359	0.8099	30.31
DQO	All	**0.7790**	**0.8873**	**0.8508**	**35.31**	**0.7212**	**0.8525**	**0.8255**	**31.21**
Baseline	$\mathcal{T}$	0.7231	0.8417	0.8272	34.50	0.6677	0.8040	0.7979	32.62
DQO	$\mathcal{T}$	**0.7381**	**0.8559**	**0.8401**	**35.34**	**0.6854**	**0.8209**	**0.8137**	**33.16**
Baseline	$\mathcal{T}^c$	0.7691	0.8805	0.8410	34.13	0.7084	0.8423	0.8123	29.85
DQO	$\mathcal{T}^c$	**0.7872**	**0.8935**	**0.8529**	**35.30**	**0.7284**	**0.8588**	**0.8278**	**30.82**
Baseline	$\mathcal{R} \cap \mathcal{T}^c$	0.7802	0.8820	0.8447	36.46	0.7202	0.8432	0.8154	33.01
DQO	$\mathcal{R} \cap \mathcal{T}^c$	**0.7967**	**0.8936**	**0.8557**	**37.54**	**0.7391**	**0.8593**	**0.8307**	**34.13**
Baseline	$\mathcal{R}^c$	0.7549	0.8787	0.8364	31.17	0.6934	0.8413	0.8084	25.84
DQO	$\mathcal{R}^c$	**0.7751**	**0.8934**	**0.8493**	**32.46**	**0.7147**	**0.8581**	**0.8242**	**26.61**

Table 1: **Evaluation metrics on FLORES+ devtest and NTREX** with the NVIDIA Megatron EN-X model, before and after task-alignment using DQO. Results are shown for relevant groupings of the 30 target languages: all languages, languages used in DQO ($\mathcal{T}$), languages not used in DQO ($\mathcal{T}^c$), languages not used in DQO, but related to those used in DQO ($\mathcal{R} \cap \mathcal{T}^c$), and languages neither used nor related to the languages used in DQO ($\mathcal{R}^c$).

and avoidance of unique language features (Koppel and Ordan, 2011; Laviosa, 1998; Tirkkonen-Condit, 2004). These effects are significant enough that classification models can distinguish translated and original text with high accuracy (Baroni and Bernardini, 2005; Sominsky and Wintner, 2019), as well as identifying the source language of the text (Koppel and Ordan, 2011). As humans show a consistent preference for translations closer to the distribution of original text rather than translationese (Riley et al., 2020; Freitag et al., 2022b), this creates an inherent task–data mismatch for training data translated in the source–target direction.

Source–Target Domain Mismatch. Translation pairs in the other direction, target–source, are better aligned with human preference, as the target labels are drawn from the original text distribution rather than from translationese. Unfortunately, they suffer from another subtle source of task–data mismatch found in human translations: Source–target domain mismatch (Shen et al., 2021) is the observation that speakers of different languages tend to discuss different topics. For instance, a Cherokee newspaper is likely to report on different topics than an Icelandic newspaper would, and translations of these topics would remain representative of the Cherokee domain. This effect is especially pronounced for low-resource language pairs (Shen et al., 2021).

If one were to avoid the task–data mismatch of translationese by using only target–source translation pairs, the training data may lack key information about topics found only in the source domain. Because the task is translation from the source domain into the target language, this, too, would represent an unavoidable task–data mismatch.

3 Human Preference Learning for LLMs

Supervised data showing chat-based dialog between humans and AI assistants was, prior to the wide availability of such agents in the form of LLMs, understandably rare. Even with the advent of high quality proprietary and open-source models, which one could sample to create synthetic data, there is a fundamental task–data mismatch: the task is not to imitate an existing AI assistant, but (ideally) to train a new state-of-the-art model.

LLM training instead follows a two-step process:

1. Supervised learning on massive amounts of web data.
2. Task alignment using instruction fine-tuning and human preference learning.

In step one, the actual task for which the model is optimized is predicting the next token in documents taken from the web. This, when done at scale and with a variety of data sources, provides the model with extensive world knowledge and understanding of a wide array of styles and document types.

This is then followed by instruction fine-tuning, a comparatively brief round of supervised learning on human- or AI-labeled examples of dialogues, which brings the model's output distribution into the general neighborhood of desired behavior. Finally human preference learning, using actual human rankings aligns the model with the desired task: producing human-preferred responses to questions and dialog, while remaining helpful and harmless (Bai et al., 2022).

Direct Preference Optimization (DPO) is a preference learning algorithm that trains on preference pairs of the form (x, y_w, y_l), with x being a model input, and y_w and y_l being two potential model outputs for the input x, marked as chosen (winning) or rejected (losing) by humans during data collection (Rafailov et al., 2023), using the loss function:

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

where σ is the logistic function.

4 Direct Quality Optimization for NMT

Because of its stability and ease of use, we select DPO as the basis for our experiments with human preference learning as a form of task alignment for NMT. As a proxy for human preferences, we use the CometKiwi22 quality estimation model to score and compare multiple translations of a given source (Rei et al., 2022b). CometKiwi22 is highly multilingual and has been shown to correlate well with human preference (Kocmi et al., 2024). To verify that our method is not dependent on the specific choice of quality estimation model we conducted a brief experiment using MetricX (Juraska et al., 2023) instead of CometKiwi22 and obtained very similar results.

Our main experiments are run with the NVIDIA Megatron English–Many model[1], a 500M parameter encoder-decoder model, which supports translating from English into 30 languages[2] from 14 language families, listed in Table 2. We denote the complete list of supported target languages as $\mathcal{M}$.

Language Family	Languages (ISO 639-1)
Baltic	lt, lv
Germanic	da, **de**, nl, no, sv
Romance	**es**, fr, it, pt, ro
Slavic	bg, cs, hr, pl, **ru**, sl, uk
Uralic	et, fi, hu
Other	el, **hi**, id, ja, ko, tr, vi, **zh**

Table 2: **Target languages supported by the NVIDIA Megatron En-X model**. The category "Other" contains all languages that are the only supported representative of their language family. The languages on which we apply task alignment are denoted in boldface.

[1] https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/megatronnmt_en_any_500m

[2] The model was originally trained to support 32 languages, but we found that translating into Arabic and Slovak resulted in degenerate output.

The model's multilingual nature allows us to apply task alignment to a subset of language pairs and observe the effects on unrelated languages, with minimal risk of exposing the model to any new information in those languages.

Any improvements in those languages must either apply to all languages (such as avoiding omissions or additions), or are language specific, and can only have come from previously unused latent knowledge acquired during supervised training.

In our experiments, we selected Chinese, German, Hindi, Russian and Spanish as the target languages used during task alignment, termed $\mathcal{T} = \{de, es, hi, ru, es\}$. Let $\mathcal{T}^C = \mathcal{M} \setminus \mathcal{T}$ be the set containing the 25 target languages not represented during task alignment, R be the set of languages related to at least one language in $\mathcal{T}$ (defined as belonging to the same language family), and $\mathcal{R}^C = \mathcal{M} \setminus \mathcal{R}$ be the languages unrelated to any of the languages used in task alignment. An overview of how many languages belong to each set is shown in Table 3.

Subset	Definition	Size
$\mathcal{T}$	Languages seen in DQO	5
$\mathcal{T}^C$	Languages not seen in DQO	25
$\mathcal{R}$	Languages related to $\mathcal{T}$	19
$\mathcal{R}^C$	Languages unrelated to $\mathcal{T}$	11

Table 3: **Target languages supported by the NVIDIA Megatron EN-X model**, categorized by their relationship with the languages selected for task alignment.

As the seed dataset from which to draw source sentences for human preference learning, we use the source side of a mixture of publicly available English–German MT datasets (see Appendix A.4).

From this dataset, we sample 8000 source segments. For each source segment, we sample a target language from $\mathcal{T}$, the languages used for task alignment, and use the current policy model to sample 64 translations into that language using combined Top-K and Top-P sampling, with $K = 40$, $P = 0.8$ (Fan et al., 2018; Holtzman et al., 2020). We also add the greedy translation for each source segment, obtaining a total of 520 000 translations.

Letting the output of the CometKiwi22 Quality Estimation (QE) model for a source x and translation y be $r_{QE}(x, y)$, we build a relation $\succ_x$ as a proxy for true human preferences:

$$y_1 \succ_x y_2 \equiv r_{QE}(x, y_1) > r_{QE}(x, y_2) + \varepsilon$$

where $\varepsilon \geq 0$ is a tolerance parameter to help miti-

gate proxy model noise. We set $\varepsilon = 0.005$.

To construct preference pairs, we then select the highest scoring translation per source segment as y_w and uniformly sample a single y_l from all remaining translation candidates that satisfy $y_w \succ y_l$ under our proxy model.

This results in slightly under 8000 preference pairs (occasionally the maximum difference in COMET22 score between a segment's highest and lowest scoring sampled translations is less than ε, in which case we do not produce a preference pair), we run DPO training with a batch size of 8192 tokens (counting source, chosen and rejected tokens), a learning rate of $1e{-}6$ and $\beta = 0.5$. See Appendix 8 for a full list of hyperparameters.

At this point, we train on the preference pairs using standard DPO for 8 epochs, after which we sample a fresh set of source segments from the seed dataset, sample translations from the policy model, create a new set of preference pairs, and begin the training again. This resampling process helps ensure that the preference pairs remain relevant to the policy model during training, and leads to substantial performance improvements. In total, we perform 5 rounds of DPO training. We call this end-to-end process Direct Quality Optimization (DQO), detailed formally in Algorithm 1.

DQO can be viewed as a batched online version of DPO, as the updates are performed on batches of data sampled from the policy model.

Algorithm 1: Direct Quality Optimization

Parameters: preference relation $\succ$, number of rounds n, epochs per round m, epoch size d, learning rate α, DPO regularization β, sampled translations per source k

Input: Source language seed dataset S, reference-free QE model r_{QE}, reference model π_{ref}

$\pi_\theta \leftarrow \pi_{\text{ref}}$;
for *round* $i = 1, 2, \ldots, n$ **do**
  $X \leftarrow$ **sample** d sentences from S;
  $P \leftarrow \varnothing$;
  foreach *source* $x \in X$ **do**
    $g \leftarrow \texttt{Greedy}_{\pi_\theta}(x)$;
    $Y \leftarrow$ **sample** k translations of x from π_θ;
    $Y_+ \leftarrow Y \cup \{g\}$;
    $y_w \leftarrow \text{argmax}_{y \in Y_+} r_{QE}(x, y)$;
    $Y_l = \{y' \in Y_+ | y_w \succ_x y'\}$;
    if $Y_l \neq \varnothing$ **then**
      $y_l \leftarrow$ **sample** $y \in Y_l$;
      $P \leftarrow P \cup \{(x, y_w, y_l)\}$;
  end
  for *epoch* $j = 1, 2, \ldots, m$ **do**
    $\pi_\theta \leftarrow \texttt{DPO}(\pi_\theta, \pi_{ref}, P, \alpha, \beta)$;
  end
end

Figure 1: **Direct Quality Optimization (DQO).** $\texttt{Greedy}_\pi(x)$ is the translation of x produced with greedy search and the model π. `DPO` refers to Direct Preference Optimization – for full implementation details see Rafailov et al. (2023).

5 Experimental Results

5.1 Automatic Quality Metrics

We evaluated the model pre- and post-task alignment on the FLORES+ (Team et al., 2024) and NTREX (Federmann et al., 2022; Barrault et al., 2019) datasets, both of which cover all of the languages supported by the Megatron model.

We use corpus-level sacreBLEU[3] (Post, 2018) as well as three neural evaluation models: Reference-free CometKiwi22 (Rei et al., 2022b), reference-based COMET22 (Rei et al., 2022a), and BLEURT (Sellam et al., 2020).

Here it is important to note that the CometKiwi22 model was used as a proxy for human preferences in this experiment, and was thus directly optimized for. The scores from the other two neural evaluation models are thus more reliable measures of general model quality, and allow us to check for reward hacking, i.e. over-optimization for the CometKiwi22 model at the cost of performance.

Results are reported in Table 1 and Figure 2. We find that DPO task alignment increases all three neural quality metrics on both datasets for each of the 30 target languages. BLEU scores increased for all languages on both datasets, with the exception of Hindi, which decreased by 0.40 BLEU on NTREX and 0.4 BLEU on FLORES+ devtest, despite showing improvements on the three neural metrics, like all other languages.

Significantly, translation quality, as measured by all four translation quality metrics, improved even

[3]Signature: `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.0`. For JA and ZH, we additionally use the `mecab-ja` and `mecab-zh` tokenizers.

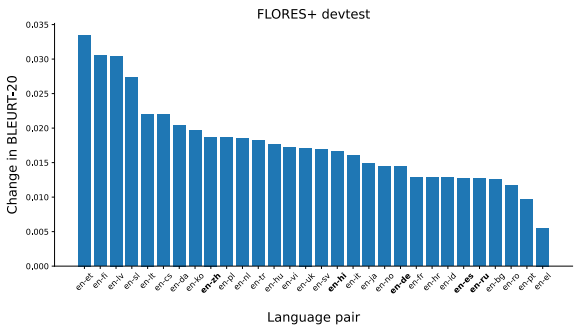

Figure 2: **Changes in BLEURT on FLORES+ devtest** with the NVIDIA Megatron EN-X model, before and after task alignment with DQO. Languages used in DQO are bolded.

for target languages unrelated to the languages used in DPO task alignment. See Appendix A.5 for the metrics for each individual language.

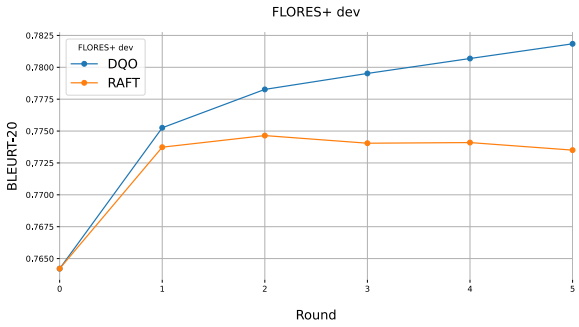

Figure 3: **Mean BLEURT-20 on FLORES+ dev at each round of DQO** with the NVIDIA Megatron EN-X model, using either Direct Quality Optimization (DQO) or Supervised Fine-Tuning (RAFT) to update the model.

To ablate the use of DPO as the update step within DQO, we perform a comparative experiment identical to DQO as described in Section 4 and Algorithm 1, but using standard supervised fine-tuning (SFT) on the preferred translation instead of DPO. Note that this is equivalent to Reward rAnked Fine-Tuning (RAFT) (Dong et al., 2023). Figure 3 shows mean performance for all language pairs through the 5 rounds of DQO. RAFT's lower performance is primarily due to catastrophic behavior for FR, JA, KO, and ZH. The poor performance of RAFT on FR, JA and KO, which were not used in RAFT training, could potentially be explained by a failure to generalize from the training data languages. Regarding ZH, which was one of the five languages used in training, we suspected unintentionally reversed labels. However, careful inspection of the training preference pairs showed no issues. See Appendix A.6, Figure 4 for charts of individual language performance and Figure 5 for mean performance after excluding the above mentioned outliers. We leave a deeper analysis to future work.

5.2 Training Data Perplexity

In order to confirm the existence of a task–data mismatch, we examine DQO's effect on model perplexity over the training data. As we do not have access to the training data used for the NVIDIA Megatron English-Many model, we repeat the above experiment with a proprietary encoder-decoder model trained on publicly available English-to-German data using the NVIDIA NeMo framework (Kuchaiev et al., 2019) (See Appendix A.4). The model architecture is similar to the Megatron model, and follows the deep encoder, shallow decoder recipe from (Kasai et al., 2021), but is larger, with a model width of 2048, a feed-forward width of 8192, 21 encoder layers, 2 decoder layers, and a 32k token vocabulary, totaling 1.3B parameters.

We apply DQO to this model as with the Megatron model, however using only English–German preference pairs. After applying DQO, we see large improvements in CometKiwi22 and COMET22 for a variety of evaluation datasets, confirming that DQO worked as expected. The arithmetic mean of perplexity over a random sample of 1 million segments from the training data increased from 7.219 (baseline model) to 9.435 (DQO), confirming that the improvements in test data preference correspond to a reduction in the model's fit to the training corpus.

5.3 Discussion

The nearly-universal improvements for both FLORES+ and NTREX in all four automatic translation quality metrics (Table 1) provide strong evidence that DQO is a suitable task-alignment algorithm for the task of producing human-preferred translations.

As shown in Section 5.2, while improving task performance, DQO increases perplexity over the training data used during supervised training. This, combined with the finding that DQO is a suitable task alignment algorithm, is evidence for the existence of the task–data mismatch.

Much of this improvement can likely be credited to general, language-agnostic changes in model behavior, even with the restriction to using only 5 of the 30 supported target languages in DQO. If task alignment of a model with a given target lan-

Source	... under the leadership of **Deng Xiaoping**.
Baseline	... tika veiktas ***Deng Xiaoping*** vadībā.
DQO	... tika veiktas **Dena Sjaopina** vadībā.
Source	... that **Carolyn Wilson** of the OHA had stolen their security deposits ...
Baseline	... ka OHA ***Carolyn Wilson*** bija nozagusi viņu drošības depozītus ...
DQO	... ka OHA darbiniece **Karolīna Vilsona** bija nozagusi viņu drošības depozītus ...
Source	... that it was **Louis Jourdain**, 16-year old son of ... **Floyd Jourdain**.
Baseline	... ka tas bija ***Louis Jourdain***, 16 gadus vecs ... **Floida *Jourdaina*** dēls.
DQO	... ka tas bija **Luiss Džordēns**, 16 gadus vecs ... **Floida Džordēna** dēls.
Source	**King Sejong** was the fourth king of the **Joseon** Dynasty ...
Baseline	***King Sejong*** bija ceturtais karalis no ***Joseon*** dinastijas ...
DQO	**Karalis Sedžons** bija ceturtais **Džosona** dinastijas karalis ...

Table 4: **Examples of translations into Latvian from the FLORES+ data set before and after DQO.** Names are bolded to highlight the DQO model's increased ability to consistently transliterate names into Latvian orthography. Names that are incorrectly transliterated are in italics. Sentences are truncated to avoid dataset leakage.

guage reduces the likelihood of untranslated source text, for instance, it would not be surprising to see similar improvements in other target languages.

Similarly, if task alignment for a given target language led to language-specific improvements (e.g., in grammar, sentence structure, punctuation, general fluency, etc.), it seems plausible that transfer learning could lead to improvements in closely related languages that have similar features.

However, manual inspection of translations before and after DQO revealed language-specific improvements in unrelated languages. In Latvian, for instance, foreign names are transliterated to match Latvian orthography and declined for grammatical case and gender, e.g. Klavinska (2021) report that *George Clooney* should be translated as *Džordžs Klūnijs*. While the baseline model applies correct transliteration occasionally and inconsistently, the DQO model almost always produces the correct transliteration. Several examples are included in Table 4.

As DQO was only performed on Chinese, German, Hindi, Russian or Spanish, none of which are closely related to Latvian, this behavior cannot have been learned from scratch during DQO. Although Chinese, Hindi, and Russian also transcribe foreign names, they use non-Latin scripts.

One possible explanation is that the baseline model learned to model both transliteration and non-transliteration, due to the range of translation quality in its supervised training data, causing inconsistent behavior at inference time. When DQO then shifts the output distribution towards certain human-preferred features, the probability of any correlated features (e.g., transliteration in Latvian), also increases.

5.4 Human Evaluation

To verify the presence of further language-specific changes for unrelated languages, we performed a human evaluation using the Multidimensional Quality Metrics framework (MQM) with professional translators (Lommel et al., 2014; Freitag et al., 2021). The translators were trained on MQM and Anthea[4], the open-source tool we used for performing MQM. We follow Freitag et al. (2021) in weighting major non-translations at 25 MQM points, other major errors at 5, and all minor errors at 1, except minor punctuation errors, which are 0.1 points.

For analysis, we selected two target languages not closely related to the languages used for task alignment: Lithuanian and Japanese.

These were selected to provide one low-to-medium resource language written in the Latin script and one in a non-Latin script, because neither is an outlier in quality metric improvement compared to the other supported language pairs, and to avoid the bias of examining Latvian, which we had already manually inspected.

For each language, we sampled complete documents (each generally two to five sentences forming a single paragraph) from FLORES+ until we had 100 source segments. The translators then annotated the baseline and task-aligned translations.

We then sorted the MQM error subcategories into two buckets, language agnostic and language specific, as seen in Table 7 in Appendix A.1.

We observe reduced error rates in both Japanese and Lithuanian in both the language-agnostic and language-specific categories (Table 5). The over-

[4] https://github.com/google-research/google-research/tree/a676d87/anthea

Language	Model	Severity				Language Specific			Weighted MQM ↓
		NT	Major	Minor	Trivial	Yes	No	N/A	
Japanese	Baseline	0	1.15	**0.61**	0.06	1.28	0.50	0.01	6.256
	DQO	0	**0.93**	0.63	**0.03**	**1.16**	**0.40**	0.01	**5.223**
Lithuanian	Baseline	0.03	0.95	0.89	0.12	1.48	0.51	0	6.402
	DQO	**0.01**	**0.80**	**0.77**	**0.10**	**1.24**	**0.44**	0	**5.030**

Table 5: **Mean number of Multidimensional Quality Metrics (MQM) errors per segment**, as annotated by professional human evaluators, with two different groupings: by severity and by whether the MQM subcategory is language specific or agnostic. NT stands for non-translation, i.e., a segment that cannot be construed as a translation of the source. Trivial refers to minor punctuation errors. This covers 100 randomly sampled English segments from the FLORES+ dataset, translated by the NVIDIA Megatron model before task alignment (baseline) and after task alignment (DQO). The weighted MQM score follows Freitag et al. (2021).

all weighted MQM score also decreased for both languages, with significant improvements in both Lithuanian ($p_u = .001$) and Japanese ($p_u = .012$), where p_u-values are conservative estimates of the true p-values computed using paired one-sided approximate randomization (Phipson and Smyth, 2010) with the Marot toolkit.[5]

5.5 DQO for Large Language Models

To compare DQO's performance against the strong baseline of other state-of-the-art DPO variants on a large language model trained specifically for translation, we apply it to the Alma-13B-LoRA model, a LLaMA-2-13B model with continued pretraining on Chinese, Czech, English, German, Icelandic, and Russian monolingual data and LoRA fine-tuning on high quality translation data (Xu et al., 2024a; Hu et al., 2022).

The highest performing human preference alignment method previously reported for this model is Contrastive Preference Optimization (CPO), a variant of DPO applied to the Alma-13B-LoRA model to create Alma-13B-R (Xu et al., 2024b). To ensure a direct comparison of optimization methods, we adopt the same data conditions and parameter masks as that prior work: restricting our seed dataset to the training data used for Alma-13B-R (the combined FLORES+ dev and devtest splits), fine-tuning only the LoRA adapters of the model, and evaluating translation out of English on the WMT'21 (for Icelandic) and WMT'22 (for the other languages) datasets.

Due to the restricted seed dataset used in this experiment, source segments are reused between rounds. As in previous experiments, we sample 8000 source segments, sample 64 translations per segment (as well as the greedy translation), and use CometKiwi22 as a proxy for human preferences. Other hyperparameters were adjusted based on a manual hyperparameter search to accommodate the differing training and sampling dynamics of LoRA training with an LLM (see Appendix A.3 for all hyperparameters).

Table 6 shows the results. The translations for ALMA-13-LoRA and ALMA-13B-R are generated with greedy inference on the publicly available model parameters[6]. This experiment indicates that DQO maintains a substantially higher BLEU score than CPO while providing similar improvements in BLEURT, COMET22, and CometKiwi22. Unlike our encoder-decoder experiments, source segments were reused between rounds to achieve a fair comparison with CPO. We would expect a higher performance with a larger pool of source data, but leave confirmation of this assumption to future work.

6 Related Work

The idea of task–data mismatch in NMT is not new. There has been extensive previous work focused on reducing this mismatch through data filtering, using surface-level heuristics (Koehn et al., 2007), statistical and neural models for alignment and quality evaluation (Sánchez-Cartagena et al., 2018; Heffernan et al., 2022; Peter et al., 2023), language identification (Lui and Baldwin, 2011; Joulin et al., 2016), or ensembles (Koehn et al., 2020).

While data filtering techniques do help reduce the task–data mismatch, they force a trade-off between increasing task alignment and retaining flawed, but potentially useful, training data. To

[5] https://github.com/google-research/google-research/tree/a676d87/marot/README.md

[6] https://huggingface.co/haoranxu/ALMA-13B-Pretrain-LoRA, https://huggingface.co/haoranxu/ALMA-13B-R

Model	English → Czech				English → German			
	BLEURT	COMET22	CometKiwi22	BLEU	BLEURT	COMET22	CometKiwi22	BLEU
ALMA-13B-LoRA	79.62	88.94	83.31	**29.33**	75.06	85.14	82.19	29.65
+ DQO	80.58	89.69	**84.46**	27.72	76.03	85.95	**83.10**	**29.72**
+ CPO (ALMA-13B-R)	**80.90**	**89.73**	84.38	24.29	**76.79**	**86.24**	82.96	26.72
Model	**English → Icelandic**				**English → Russian**			
	BLEURT	COMET22	CometKiwi22	BLEU	BLEURT	COMET22	CometKiwi22	BLEU
ALMA-13B-LoRA	71.64	85.32	80.84	25.06	74.25	86.90	82.55	**27.48**
+ DQO	**72.00**	85.57	**81.72**	**25.09**	75.40	87.71	**83.68**	26.68
+ CPO (ALMA-13B-R)	71.71	**86.25**	81.20	21.03	**75.74**	**88.05**	83.63	23.12
Model	**English → Chinese (simpl.)**				**Average**			
	BLEURT	COMET22	CometKiwi22	BLEU	BLEURT	COMET22	CometKiwi22	BLEU
ALMA-13B-LoRA	69.79	85.54	80.56	**37.80**	74.07	86.37	81.89	**29.86**
+ DQO	**70.60**	**86.37**	**81.84**	35.58	74.92	87.06	**82.96**	28.96
+ CPO (ALMA-13B-R)	**70.60**	86.35	81.79	32.15	**75.15**	**87.32**	82.79	25.46

Table 6: Evaluation of DQO and CPO (Xu et al., 2024b) on the ALMA-13B-LoRA model. Scores are reported on the WMT'21 (Icelandic) and WMT'22 (remaining languages) test sets. The hyperparameters are specified in Appendix A.3.

counter this, curriculum learning can be used, by training first on a conservatively filtered dataset, then shifting to a cleaner subset of the data (Bogoychev et al., 2023).

However, no amount of data filtering can remove the effects of translationese, as it is present in all translations. Riley et al. (2020) and Freitag et al. (2022b) both address this by treating original and translated text as separate languages in a "multilingual" NMT model, by training either a classifier or a contrastive language model to tag each source and target segment as either original or translated. At inference time, they use their model in a zero-shot setting to translate from original source text into the distribution of original target text.

Similarly, Tomani et al. (2024) label each source sentence with a binned QE score. By adding the label of the highest quality bin to a source sentence at inference time, they successfully bias the model towards high quality translations.

Ramos et al. (2024) apply RLHF (Ziegler et al., 2020) to NMT using various QE metrics as reward, and compare it to data filtering, re-ranking using a QE model, and Minimum Bayes Risk decoding (MBR) (Kumar and Byrne, 2004; Freitag et al., 2022a), finding that a combination of data filtering, RLHF, and re-ranking performs best.

In DPO MBR fine-tuning, MBR is used to generate preference pairs for use with DPO (Yang et al., 2024). Compared to DQO, this method is computationally more expensive, and requires a reference-based QE model. In addition, DQO's online nature ensures that preference pairs remain relevant to the policy model.

Xu et al. (2024c) apply RLHF with a reward model trained to distinguish high quality references (from literary translations) and translations sampled from their model. Similar to us, they find evidence of cross-lingual transfer learning during preference learning. Specifically, when optimized only on EN–ZH, their model improved for EN to FR, ES, RU, and AR. When training only on EN–AR, however, they saw improvements in only half of the target languages.

Reward rAnked Fine-Tuning (RAFT) is the method most similar to DQO, but uses SFT to update the model towards a single preferred output rather than using DPO with a preferred/rejected output pair (Dong et al., 2023). As it was not evaluated for the translation task, used an independently trained reward model, and had slight differences in sampling parameters, we ran an ablation on whether to use DPO or SFT in DQO (see Section 4).

7 Conclusion

We demonstrate the existence of a fundamental task–data mismatch in NMT and introduce Direct Quality Optimization (DQO), a method of aligning pretrained models with human preference.

Using DQO on a multilingual NMT model, we find improvements in automatic quality metrics for all supported target languages, even those neither

used for DQO, nor related to the languages used for DQO. A human evaluation confirms that these improvements reflect increased human preference.

The improvements in translation quality for unrelated languages include language specific features that were not seen during DQO, suggesting that the baseline model had, but did not use, knowledge of those features during inference. We suggest that this is the expected behavior of a model trained with supervised learning, and present DQO as an efficient method of aligning a translation model with human preference.

In an experiment on ALMA-13B-LoRA we confirm that DQO is applicable to decoder-only LLMs.

8 Limitations

This work only tests one quality evaluation model as a proxy for human preferences, CometKiwi22, and does not examine the impact of that proxy's quality. We focused primarily on a single translation model, the NVIDIA Megatron English-Many model, using a 1.3B paramter English-German model only for the perplexity experiments (as we had access to the training data), and ALMA-13B-LoRA to verify applicability on decoder-only models. Human evaluation of translation quality was only performed on two language pairs. For all others, we relied on automatic quality evaluation metrics such as BLEURT, COMET22 and BLEU, which may not fully capture true human preference.

References

A.H. Albir. 2017. *Researching Translation Competence by PACTE Group*. Benjamins Translation Library. John Benjamins Publishing Company.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv:2204.05862.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Marco Baroni and Silvia Bernardini. 2005. A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text. *Literary and Linguistic Computing*, 21(3):259–274.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Nikolay Bogoychev, Jelmer van der Linde, Graeme Nail, Barry Haddow, Jaume Zaragoza-Bernabeu, Gema Ramírez-Sánchez, Lukas Weymann, Tudor Nicolae Mateiu, Jindřich Helcl, and Mikko Aulamo. 2023. OpusCleaner and OpusTrainer, open source toolkits for training machine translation and large language models. arXiv:2311.14838.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the Bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. Raft: Reward ranked finetuning for generative foundation model alignment. arXiv:2304.06767.

Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from united nation documents. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Ahmed El-Kishky, Adithya Renduchintala, James Cross, Francisco Guzmán, and Philipp Koehn. 2021. Xlent: Mining a large cross-lingual entity dataset with lexical-semantic-phonetic word alignment. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10424–10430.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022a. High Quality Rather than High Model Probability: Minimum Bayes Risk Decoding with Neural Metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Markus Freitag, David Vilar, David Grangier, Colin Cherry, and George Foster. 2022b. A natural diet: Towards improving naturalness of machine translation output. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3340–3353, Dublin, Ireland. Association for Computational Linguistics.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Marcin Junczys-Dowmunt, Bruno Pouliquen, and Christophe Mazenc. 2016. Coppa v2.0: Corpus of parallel patent applications. building large parallel corpora with gnu make.

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.

Jungo Kasai, Nikolaos Pappas, Hao Peng, James Cross, and Noah Smith. 2021. Deep encoder, shallow decoder: Reevaluating non-autoregressive machine translation. In *International Conference on Learning Representations*.

Antra Klavinska. 2021. Transcription of foreign personal names in the written works of learners of latvian as a foreign language. *Journal of Education Culture and Society*, 12:469–481.

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, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1999–2014, Bangkok, Thailand. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742, Online. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In

Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, Patrice Castonguay, Mariya Popova, Jocelyn Huang, and Jonathan M. Cohen. 2019. Nemo: a toolkit for building ai applications using neural modules. arXiv:1909.09577.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Massimo La Morgia, Alessandro Mei, Eugenio Nerio Nemmi, Luca Sabatini, and Francesco Sassi. 2023. Translated texts under the lens: From machine translation detection to source language identification. In *Advances in Intelligent Data Analysis XXI*, pages 222–235, Cham. Springer Nature Switzerland.

Sara Laviosa. 1998. Core patterns of lexical use in a comparable corpus of english narrative prose. *Meta*, 43(4):557–570.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Arle Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0:455–463.

Marco Lui and Timothy Baldwin. 2011. Cross-domain feature selection for language identification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 553–561, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Yan Meng, Di Wu, and Christof Monz. 2024. How to learn in a noisy world? self-correcting the real-world data noise on machine translation. arXiv:2407.02208.

Jan-Thorsten Peter, David Vilar, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, and Markus Freitag. 2023. There's no data like better data: Using QE metrics for MT data filtering. In *Proceedings of the Eighth Conference on Machine Translation*, pages 561–577, Singapore. Association for Computational Linguistics.

Belinda Phipson and Gordon K Smyth. 2010. Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1).

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

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 *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz-Rojas. 2020. Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal. European Association for Machine Translation.

Miguel Moura Ramos, Patrick Fernandes, António Farinhas, and André F. T. Martins. 2024. Aligning neural machine translation models: Human feedback in training and inference. arXiv:2311.09132.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova,

Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a language in "multilingual" NMT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7737–7746, Online. Association for Computational Linguistics.

Roberts Rozis and Raivis Skadiņš. 2017. Tilde MODEL - multilingual open data for EU languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 263–265, Gothenburg, Sweden. Association for Computational Linguistics.

Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez-Sánchez. 2018. Prompsit's submission to wmt 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Brussels, Belgium. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Jiajun Shen, Peng-Jen Chen, Matthew Le, Junxian He, Jiatao Gu, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2021. The source-target domain mismatch problem in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1519–1533, Online. Association for Computational Linguistics.

Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the Common Crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1374–1383, Sofia, Bulgaria. Association for Computational Linguistics.

Ilia Sominsky and Shuly Wintner. 2019. Automatic detection of translation direction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1131–1140, Varna, Bulgaria. INCOMA Ltd.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Dániel Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. *CoRR*, abs/cs/0609058.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. No Language Left Behind: Scaling neural machine translation to 200 languages. *Nature*, 630:841–846.

Brian Thompson, Mehak Dhaliwal, Peter Frisch, Tobias Domhan, and Marcello Federico. 2024. A shocking amount of the web is machine translated: Insights from multi-way parallelism. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1763–1775, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Sonja Tirkkonen-Condit. 2004. Unique items — over- or under-represented in translated language? In *Translation Universals: Do they exist?*, pages 177–184. Benjamins Translation Library.

Christian Tomani, David Vilar, Markus Freitag, Colin Cherry, Subhajit Naskar, Mara Finkelstein, Xavier Garcia, and Daniel Cremers. 2024. Quality-aware translation models: Efficient generation and quality estimation in a single model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15660–15679, Bangkok, Thailand. Association for Computational Linguistics.

Philip Williams and Barry Haddow. 2021. The elitr eca corpus. arXiv:2109.07351.

Krzysztof Wołk and Krzysztof Marasek. 2014. Building subject-aligned comparable corpora and mining it for truly parallel sentence pairs. *Procedia Technology*, 18:126–132. International workshop on Innovations in Information and Communication Science and Technology, IICST 2014, 3-5 September 2014, Warsaw, Poland.

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.

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*, ICML'24. JMLR.org.

Nuo Xu, Jun Zhao, Can Zu, Sixian Li, Lu Chen, Zhihao Zhang, Rui Zheng, Shihan Dou, Wenjuan Qin, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024c. Advancing translation preference modeling with rlhf: A step towards cost-effective solution.

Guangyu Yang, Jinghong Chen, Weizhe Lin, and Bill Byrne. 2024. Direct preference optimization for neural machine translation with minimum Bayes risk decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 391–398, Mexico City, Mexico. Association for Computational Linguistics.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences. arXiv:1909.08593.

A Appendix

A.1 MQM Error Subcategories by Generality

Language-agnostic	**Language-specific**	**Other**
Accuracy/Creative Reinterpretation	Fluency/Grammar	Other
Accuracy/Mistranslation	Fluency/Register	Source issue
Accuracy/Source language fragment	Fluency/Spelling	
Accuracy/Addition	Fluency/Punctuation	
Accuracy/Omission	Fluency/Character encoding	
Fluency/Inconsistency	Style/Unnatural or awkward	
Terminology/Inconsistent	Style/Bad sentence structure	
Non-translation	Terminology/Inappropriate for context	
	Locale convention/Address format	
	Locale convention/Date format	
	Locale convention/Currency format	
	Locale convention/Telephone format	
	Locale convention/Time format	
	Locale convention/Name format	

Table 7: **Multidimensional Quality Metrics error subcategories by generality**. *Language-agnostic errors* are those governed by a principle that can be generalized to all language pairs, e.g., that translations should not omit information. *Language-specific errors* are those that require additional, language-specific information to generalize from one language pair to another, e.g., correcting improper sentence structure requires knowledge of correct vs. incorrect sentence structures for a given language. *Other errors* cannot be assigned to either category.

A.2 Hyperparameters Used in Experiments on NVIDIA Megatron

Hyperparameter	Definition	Value
r_{QE}	Human preference proxy model	`CometKiwi22`
n	Number of rounds	5
m	Epochs per round	8
d	Epoch size (source sentences)	8000
α	Learning rate	1×10^{-6}
β	DPO regularization factor	0.5
k	Sampled translations per source	64
K	Top-K sampling parameter	40
P	Top-P sampling parameter	0.8
ε	Preference margin	0.005
–	Batch size	8096
–	Learning rate schedule	Linear with warmup
–	Learning rate warmup steps	150
–	Gradient clipping threshold (norm)	10

Table 8: **A list of all hyperparameters** used for Direct Quality Optimization in this paper's experiments.

A.3 Hyperparameters for experiments on ALMA-13B-LoRA

Hyperparameter	Definition	Value
r_{QE}	Human preference proxy model	`CometKiwi22`
n	Number of rounds	9
m	Epochs per round	4
d	Epoch size (source sentences)	8000
α	Learning rate	5×10^{-5}
β	DPO regularization factor	0.5
k	Sampled translations per source	64
K	Top-K sampling parameter	∞
P	Top-P sampling parameter	1.0
ε	Preference margin	0.005
–	Batch size	8096
–	Learning rate schedule	Linear with warmup
–	Learning rate warmup steps	150
–	Gradient clipping threshold (norm)	10

Table 9: **A list of all hyperparameters** used for Direct Quality Optimization in the experiments on ALMA-13B-LoRA. See https://github.com/lilt/dqo/blob/main/configs/alma-13b-lora-comparison-with-cpo-4.yaml

A.4 Composition of the DQO Seed Dataset

As described in Figure 1, Direct Quality Optimization requires a seed dataset containing input samples in the source language. This dataset does not need to include references, as the policy model π_θ is used to produce a diverse set of hypotheses, which are then scored under a QE model and transformed into preference pairs.

For our experiments, we used a general and varied seed dataset consisting of the English side of the following publicly available English–German datasets provided by the OPUS project (Tiedemann, 2012):

- bible-uedin (Christodouloupoulos and Steedman, 2015)
- CCAligned (El-Kishky et al., 2020)
- CCMatrix (Schwenk et al., 2021b; Fan et al., 2021)
- DGT v2019[7]
- EBC
- ELRA-W0143[8]
- ELRA-W0201
- ELRC-CORDIS_News[9]
- ELRC-CORDIS_Results[10]

[7] https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory. The European Commission retains ownership of the data.

[8] https://www.elrc-share.eu

[9] https://elrc-share.eu/repository/browse/english-french-parallel-corpus-from-cordis-project-news/e4597da00ae511e9b7d400155d026706c248250ecee54d19bef388d2a42e6d93/

[10] https://elrc-share.eu/repository/browse/german-english-parallel-corpus-from-cordis-project-results-in-brief/e70e0b920ae511e9b7d400155d026706b079d7cd7f984a98ab96380f6215f358/

- ELRC-EMEA[11]
- ELRC-EU_publications[12]
- ELRC-EUR_LEX[13]
- ELRC-Information_Portal[14]
- ELRC-presscorner_covid[15]
- EMEA
- EUBookshop
- EUConst
- EuroPat[16]
- GlobalVoices
- GNOME
- JRC-Acquis v3.0 (Steinberger et al., 2006)[17]
- KDE4
- LinguaTools-WikiTitles
- MultiUN (Eisele and Chen, 2010)
- News-Commentary (Kocmi et al., 2023)
- OpenSubtitles (Lison and Tiedemann, 2016)
- ParaCrawl (Bañón et al., 2020)
- PHP
- Tatoeba
- Tilde EESC (Rozis and Skadiņš, 2017)
- TildeMODEL (Rozis and Skadiņš, 2017)
- WikiMatrix (Schwenk et al., 2021a)
- wikimedia[18]

[11] https://elrc-share.eu/repository/browse/bilingual-corpus-made-out-of-pdf-documents-from-the-european-medicines-agency-emea-httpswwwemaeuropaeu-february-2020-en-de/d6ce198a862611ea913100155d0267064011b731322946a6b897cf495fb6f023/. This dataset has been generated out of public content available through European Medicines Agency: https://www.ema.europa.eu/, in February 2020.

[12] This dataset was generated from public content available through the Publications Office of the European Union (OP Portal), https://op.europa.eu/en/home

[13] https://elrc-share.eu/repository/browse/covid-19-eur-lex-dataset-ilingual-en-mt/cf57fe82c5af11ea913100155d026706b5596d3f449a456f983bbb4e23de81a4/

[14] https://elrc-share.eu/repository/browse/information-portal-of-the-czech-president-and-czech-castle/2c11868e088b11e6b68800155d020502c402eaf049834da0bbb019049e42098c/

[15] https://elrc-share.eu/repository/browse/covid-19-eu-presscorner-v1-dataset-bilingual-en-de/67c1519c969311ea913100155d0267063c11069dcb104114901b3160c9f7618c/

[16] https://europat.net/

[17] https://joint-research-centre.ec.europa.eu/language-technology-resources/jrc-acquis_en. The European Commission retains ownership of the data.

[18] https://dumps.wikimedia.org/other/contenttranslation/

- Wikipedia (Wołk and Marasek, 2014)
- Wikititles (Kocmi et al., 2023)
- XLEnt (El-Kishky et al., 2021)

As well as the following publicly available datasets which were not obtained through OPUS:

- ELITR ECA (Williams and Haddow, 2021)
- Europarl (Koehn, 2005)
- Tilde EMA (Rozis and Skadiņš, 2017)
- Tilde RAPID 2019 (Rozis and Skadiņš, 2017)
- WIPO COPPA (Junczys-Dowmunt et al., 2016)
- WMT13 CommonCrawl (Smith et al., 2013)

These datasets were also used to train the model used in Section 5.2.

A.5 Results by Target Language

Model	Lang.	FLORES+ devtest				NTREX			
		BLEURT	COMET22	CometKiwi22	BLEU	BLEURT	COMET22	CometKiwi22	BLEU
Baseline	bg	0.8400	0.8974	0.8524	41.80	0.7713	0.8520	0.8242	32.00
DQO	bg	**0.8526**	**0.9067**	**0.8614**	**42.70**	**0.7865**	**0.8638**	**0.8341**	**32.40**
Baseline	cs	0.7758	0.8826	0.8327	32.60	0.7282	0.8509	0.8065	30.10
DQO	cs	**0.7978**	**0.9002**	**0.8504**	**34.00**	**0.7506**	**0.8696**	**0.8255**	**30.70**
Baseline	da	0.7744	0.8942	0.8396	46.40	0.7136	0.8541	0.8145	37.40
DQO	da	**0.7948**	**0.9091**	**0.8565**	**48.60**	**0.7355**	**0.8721**	**0.8341**	**39.30**
Baseline	de	0.7417	0.8535	0.8222	38.80	0.6793	0.8100	0.7950	30.80
DQO	de	**0.7561**	**0.8682**	**0.8338**	**39.30**	**0.7041**	**0.8315**	**0.8117**	**31.80**
Baseline	el	0.6738	0.8641	0.8032	25.90	0.6477	0.8494	0.7876	30.60
DQO	el	**0.6793**	**0.8699**	**0.8044**	**26.60**	**0.6567**	**0.8585**	**0.7892**	**31.60**
Baseline	es	0.7467	0.8567	0.8569	27.50	0.7304	0.8474	0.8330	40.50
DQO	es	**0.7594**	**0.8656**	**0.8662**	**28.80**	**0.7421**	**0.8547**	**0.8425**	**41.00**
Baseline	et	0.7779	0.8792	0.8421	27.10	0.7279	0.8451	0.8155	24.20
DQO	et	**0.8114**	**0.9041**	**0.8647**	**28.90**	**0.7603**	**0.8690**	**0.8399**	**25.00**
Baseline	fi	0.7959	0.8899	0.8471	24.40	0.7393	0.8550	0.8247	18.70
DQO	fi	**0.8264**	**0.9105**	**0.8640**	**26.00**	**0.7640**	**0.8736**	**0.8421**	**19.60**
Baseline	fr	0.7400	0.8638	0.8486	49.40	0.6525	0.8221	0.8289	36.10
DQO	fr	**0.7529**	**0.8713**	**0.8544**	**50.70**	**0.6632**	**0.8305**	**0.8344**	**37.00**
Baseline	hi	0.6825	0.7645	0.8040	**32.90**	0.6313	0.7227	0.7735	**25.50**
DQO	hi	**0.6991**	**0.7862**	**0.8217**	32.50	**0.6511**	**0.7459**	**0.7972**	25.10
Baseline	hr	0.8190	0.8942	0.8624	31.10	0.7707	0.8644	0.8326	31.80
DQO	hr	**0.8318**	**0.9032**	**0.8695**	**32.10**	**0.7847**	**0.8770**	**0.8445**	**32.50**
Baseline	hu	0.8378	0.8645	0.8354	26.90	0.7616	0.8141	0.8118	17.40
DQO	hu	**0.8554**	**0.8800**	**0.8488**	**27.10**	**0.7793**	**0.8294**	**0.8268**	**18.00**
Baseline	id	0.8030	0.9092	0.8414	47.50	0.7648	0.8823	0.8111	40.50
DQO	id	**0.8158**	**0.9172**	**0.8516**	**49.30**	**0.7784**	**0.8917**	**0.8251**	**41.10**
Baseline	it	0.7699	0.8725	0.8590	30.60	0.7280	0.8455	0.8279	36.70
DQO	it	**0.7860**	**0.8821**	**0.8676**	**31.40**	**0.7467**	**0.8613**	**0.8434**	**37.50**
Baseline	ja	0.6832	0.8918	0.8545	32.60	0.6042	0.8584	0.8251	26.40
DQO	ja	**0.6981**	**0.9019**	**0.8629**	**34.10**	**0.6208**	**0.8713**	**0.8395**	**27.10**
Baseline	ko	0.6538	0.8689	0.8433	29.40	0.5788	0.8317	0.8085	25.50
DQO	ko	**0.6734**	**0.8820**	**0.8550**	**30.30**	**0.5980**	**0.8481**	**0.8250**	**26.50**
Baseline	lt	0.8043	0.8742	0.8344	27.30	0.7485	0.8404	0.8057	21.60
DQO	lt	**0.8264**	**0.8910**	**0.8490**	**28.80**	**0.7699**	**0.8564**	**0.8181**	**22.30**
Baseline	lv	0.7896	0.8677	0.8253	30.50	0.6997	0.8097	0.7816	20.40
DQO	lv	**0.8201**	**0.8902**	**0.8431**	**32.10**	**0.7418**	**0.8424**	**0.8088**	**21.70**
Baseline	nl	0.7425	0.8617	0.8483	27.00	0.7080	0.8384	0.8205	34.20
DQO	nl	**0.7611**	**0.8756**	**0.8601**	**28.10**	**0.7262**	**0.8556**	**0.8356**	**35.40**
Baseline	no	0.7771	0.8899	0.8526	33.80	0.7447	0.8622	0.8267	36.90
DQO	no	**0.7915**	**0.8991**	**0.8646**	**34.00**	**0.7644**	**0.8779**	**0.8445**	**38.70**
Baseline	pl	0.7600	0.8678	0.8206	21.40	0.6992	0.8312	0.7939	25.70
DQO	pl	**0.7787**	**0.8818**	**0.8312**	**22.80**	**0.7153**	**0.8463**	**0.8058**	**26.80**
Baseline	pt	0.7856	0.8941	0.8453	50.80	0.7069	0.8477	0.8236	33.90
DQO	pt	**0.7952**	**0.9000**	**0.8531**	**51.20**	**0.7197**	**0.8574**	**0.8341**	**35.00**
Baseline	ro	0.8026	0.8927	0.8594	40.30	0.7338	0.8441	0.8255	33.30
DQO	ro	**0.8144**	**0.9015**	**0.8645**	**41.40**	**0.7474**	**0.8571**	**0.8386**	**34.70**
Baseline	ru	0.7430	0.8755	0.8329	31.30	0.6706	0.8299	0.8002	31.80
DQO	ru	**0.7556**	**0.8842**	**0.8419**	**32.00**	**0.6831**	**0.8433**	**0.8104**	**31.90**
Baseline	sl	0.7978	0.8679	0.8359	30.00	0.7174	0.8106	0.7877	28.30
DQO	sl	**0.8252**	**0.8860**	**0.8517**	**31.80**	**0.7576**	**0.8410**	**0.8163**	**29.60**
Baseline	sv	0.7945	0.8957	0.8515	45.40	0.7401	0.8581	0.8192	40.90
DQO	sv	**0.8113**	**0.9064**	**0.8650**	**46.20**	**0.7632**	**0.8781**	**0.8400**	**42.40**
Baseline	tr	0.7693	0.8827	0.8441	29.10	0.6802	0.8235	0.8129	17.60
DQO	tr	**0.7875**	**0.8953**	**0.8559**	**30.10**	**0.7011**	**0.8402**	**0.8287**	**17.70**
Baseline	uk	0.7432	0.8728	0.8172	29.80	0.6678	0.8230	0.7838	24.80
DQO	uk	**0.7603**	**0.8878**	**0.8300**	**30.50**	**0.6868**	**0.8423**	**0.7983**	**25.80**
Baseline	vi	0.7157	0.8736	0.8299	42.20	0.6753	0.8442	0.8081	41.30
DQO	vi	**0.7329**	**0.8857**	**0.8429**	**43.80**	**0.6917**	**0.8589**	**0.8234**	**42.10**
Baseline	zh	0.7015	0.8582	0.8199	42.00	0.6267	0.8099	0.7879	34.50
DQO	zh	**0.7202**	**0.8752**	**0.8367**	**44.10**	**0.6468**	**0.8292**	**0.8067**	**36.00**

Table 10: Automatic quality evaluation metrics for all target languages supported by the NVIDIA Megatron model, before and after Direct Quality Optimization (DQO), computed on both the FLORES+ devtest and NTREX datasets.

A.6 Ablation of Update Step: DPO vs. SFT

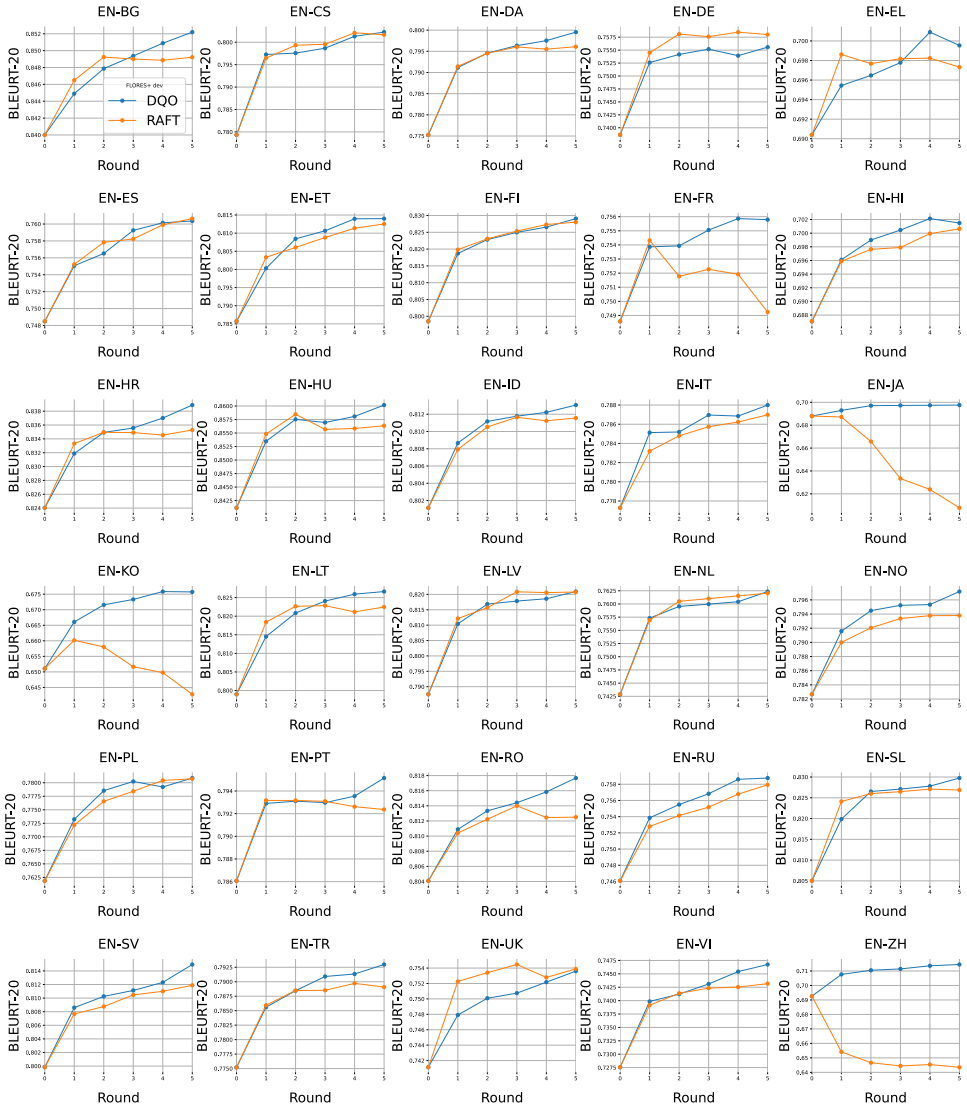

Figure 4: **Mean BLEURT-20 per language pair on FLORES+ dev after each round of DQO** with the NVIDIA Megatron EN-X model, using either Direct Preference Optimization (DPO) or Supervised Fine-Tuning (SFT) to update the model. DQO with SFT is equivalent to Reward rAnked Fine-Tuning (RAFT).

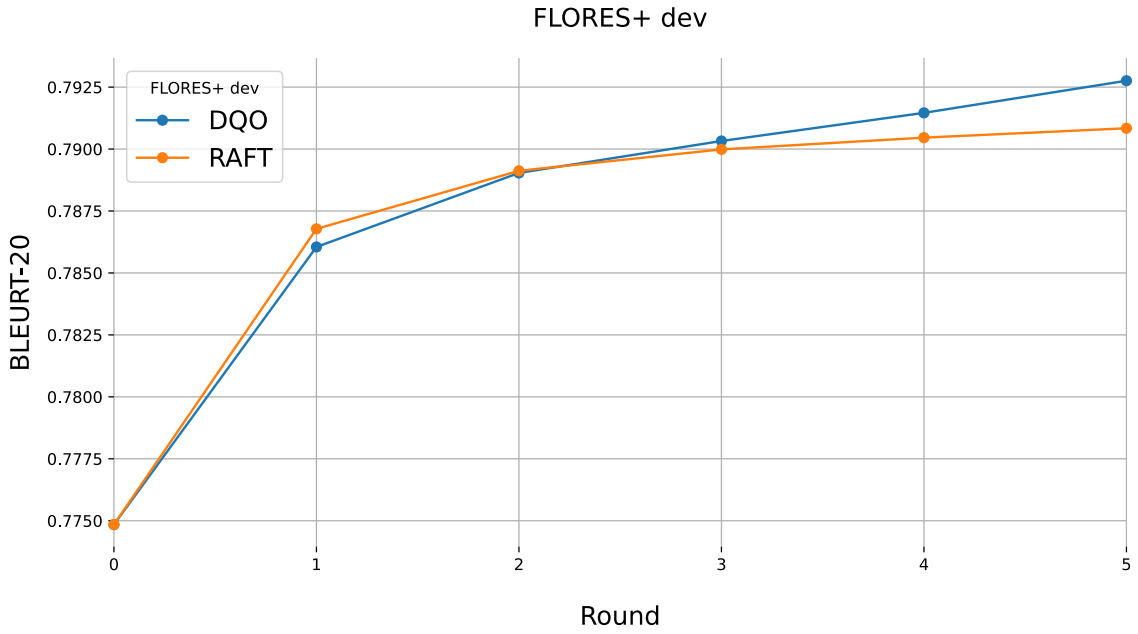

Figure 5: **Mean BLEURT-20 on FLORES+ dev, excluding outliers after each round of DQO** with the NVIDIA Megatron EN-X model, using either Direct Preference Optimization (DPO) or Supervised Fine-Tuning (SFT) to update the model. DQO with SFT is equivalent to Reward rAnked Fine-Tuning (RAFT). English to French, Chinese, Japanese, and Korean were excluded from this chart as outliers. See Figure 3 for the chart including outliers.