Mending the Holes: Mitigating Reward Hacking in Reinforcement Learning for Multilingual LLMs

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have demonstrated remarkable capability in machine translation on high-resource language pairs, yet their performance on low-resource translation still lags behind. Existing post-training methods rely heavily on high-quality parallel data, which are often scarce or unavailable for low-resource languages. In this paper, we introduce **WALAR**, a reinforcement training method using only monolingual text to elevate LLMs' translation capabilities on massive low-resource languages while retaining their performance on high-resource languages. Our key insight is based on the observation of failure modes (or "holes") in existing source-based multilingual quality estimation (QE) models. Reinforcement learning (RL) using these QE models tends to amplify such holes, resulting in poorer multilingual LLMs. We develop techniques including word alignment and language alignment to mitigate such holes in **WALAR**'s reward for RL training. We continually trained an LLM supporting translation of 101 languages using **WALAR**. The experiments show that our new model outperforms LLaMAX, one of the strongest open-source multilingual LLMs by a large margin on 1400 language directions on FLORES-101 dataset.

1 Introduction

Large Language Models (LLMs) exhibit strong capability on language translation, especially on high-resource language directions (Brown et al., 2020; Ouyang et al., 2022; Touvron et al., 2023; Zhu et al., 2024). Recent progress in open source LLMs continuously pushes the quality of machine translation to a new level on par with human (Rei et al., 2025; Grattafiori et al., 2024; Yang et al., 2025). However, their translation performance on low-resource languages remains markedly inferior.(Zhu et al., 2024; Ochieng et al., 2025). Prior works on improving LLMs' translation capabilities focus primarily on post-training strategies such as supervised fine-tuning, knowledge distillation, and back-translation (Li et al., 2024; Cheng et al., 2025). Despite the advancements, these methods are far from effective for low-resource or zero-resource languages since they rely on large amounts of high-quality parallel or preference data, which are scarce or unavailable for those languages.

Source (English): Dr. Ehud Ur, professor of medicine at Dalhousie University in Halifax, Nova Scotia and chair of the clinical and scientific division of the Canadian Diabetes Association cautioned that the research is still in its early days.

Reference (Javanese): Dr. Ehud Ur, profesor ilmu kedokteran ing Universitas Dalhousie ing Halifax, Nova Scotia lan ketua divisi klinis lan ilmiah saka Asosiasi Diabetes Kanada ngengetake menawa panaliten iku isih ing tahap wiwitan.

Over-translation (Javanese): Dr. Ehud Ur, Ketua Bagian Klinis lan Riset Asosiasi Diabetes Kanada lan profesor ing Sekolah Kedokteran Halifax Universitas Dalhousie, nedahake manawa riset iki isih ana ing tahap awal lan mbutuhake penyelidikan sing luwih jero. MetricX: -2.00

Wrong Language (Indonesian): Dr. Ehud Ur, dosen kedokteran di Universitas Dalhousie di Halifax, Nova Scotia, dan ketua divisi klinis dan ilmiah Asosiasi Diabetes Kanada memperingatkan bahwa penelitian ini masih dalam tahap awal. MetricX: -0.85

Figure 1: A case from English to Javanese showing weakness of source-based quality estimate metric (MetricX, higher is better). Even though there is over-translation and wrong candidate language, MetricX still scores high (-5 or below is considered a major error). RL using such a reward would amplify LLM's failure.

We consider the following problem: can we effectively post-train an LLM with only monolingual data to improve translation performance on massive languages? Reinforcement learning (RL) has been applied effectively to improve standalone machine translation models and LLMs (Kumar et al., 2019; Yan et al., 2023; He et al., 2024; Ramos et al., 2024). The general idea is using a metric model such as COMET (Rei et al., 2020) or COMET-Kiwi (Rei et al., 2022) to provide reward signals during RL training. The former is reference-based — compar-

ing LLM's generation candidates to references — while the latter is source-based. Since our scenario only contains monolingual text from multiple languages, we are forced to use source-based quality estimation (QE) models (Rei et al., 2022; Juraska et al., 2024).

However, directly applying RL on LLMs with quality-estimation rewards presents notable weaknesses. Our study shows that, although state-of-the-art quality estimation models achieve strong performance in evaluating translation quality (Freitag et al., 2024), these QEs exhibit noticeable holes when applied to LLM training, such as failure to detect over- and under-translation, and wrong language words. Figure 1 illustrates examples of MetricX's inability to score major translation errors. Even worse, when trained with such QE rewards, an LLM could amplify holes in certain language directions, leading to reward hacking and resulting LLM just repeating input source sentences. Astonishingly, an QE model will give a perfect score to the generated repeating source when compared to the source utterance.

To solve this major challenge, we develop **WALAR**, an effective reinforcement learning method using monolingual-only data to enhance a pre-trained LLM's multilingual translation performance. Our key idea is to use a source-based quality estimation model as the base RL reward and to mitigate its holes with additional word alignment and language alignment scores. Word alignment will encourage proper coverage, not too many left or extra words in the candidate, compared to the source utterance. Language alignment will ensure the model is generating desired target languages. We integrate all these three components in group relative policy optimization (GRPO) training framework and post-train an LLM based on LLaMAX-8B (Lu et al., 2024). The outcome and our contributions are as follows:

- We discover holes (failure modes) in widely-adopted QE models (xCOMET, MetricX) and observe that LLMs trained with these QEs lead to reward hacking in translating certain languages.
- We develop **WALAR**, a reinforcement learning method for post-training multilingual LLM with a hybrid reward to mitigate reward hacking.
- We trained an 8B LLM using our **WALAR**. Our experiments demonstrate that our model outperforms the strongest prior LLM of the same size in 1,400 language directions on the FLORES-101 dataset. Furthermore, **WALAR** generalizes across languages, improving the quality of multilingual translation even for unseen language directions during training.

2 Related Work

Reinforcement Learning in Machine Translation Performing RL on a machine translation task is not a novel idea. Feng et al. (2025) employs a reference-based model as the reward in the reinforcement learning to incorporate reasoning into LLMs' translating behavior. Ramos et al. (2025) leverages xCOMET as the reward model to generate token-level rewards, thus bringing a more fine-grained feedback and offering more benefit over sentence-level feedback. However, these works rely heavily on reference translation data. Other efforts have investigated the use of QE models in this context. Ramos et al. (2024) explores the potential of using the QE model as a data filter, reward model, and decoding reranker, demonstrating notable improvements in translation quality, whereas He et al. (2024) adopts QE-based feedback training and introduces heuristic rules to penalize the overoptimization problem of QE models. Despite the promising results achieved by previous work, whether reinforcement learning can enhance LLMs' translation abilities in low-resource languages remains important and underexplored. Our work further unleashes the power of reinforcement learning for machine translation and enhances LLMs' translation capabilities across more than 100 languages.

Multilingual LLMs Recent progress in LLMs has continuously increased the supporting language numbers of LLMs (Yang et al., 2025; Grattafiori et al., 2024; Xu et al., 2025) and achieved promising results on high-resource languages (Rei et al., 2025; Cheng et al., 2025). But the performance gap between high- and low-resource languages remains significant (Yuan et al., 2024; Zhu et al., 2024). Efforts to address such a gap either focus on the pre-training phase (Lu et al., 2024) or the post-training phase (Rei et al., 2025; Cheng et al., 2025). However, post-training methods, including instruction tuning and preference optimization, fail short in low-resource languages due to the scarcity of high-quality parallel data (Tran et al., 2020; Dang et al., 2024a). **WALAR** offers promising potential to address this problem by utilizing the abundant monolingual data in low-resource languages,

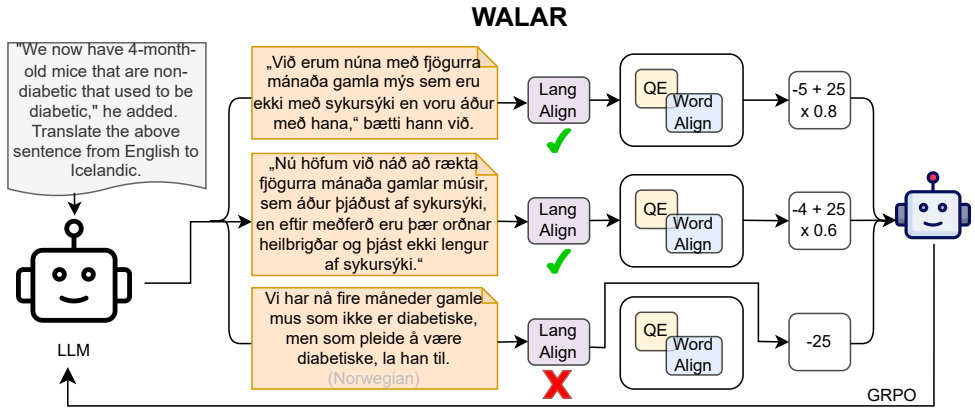

Figure 2: Illustration of **WALAR**. On each step, the LLM is prompted to translate one monolingual sentence into another language with several different rollouts. Each output will then be evaluated by language alignment, quality estimation, and word alignment. Finally, the LLM is trained using GRPO with the reward on the previous step iteratively.

thereby incentivizing LLMs' translation capabilities solely with monolingual data.

3 Proposed WALAR Method

In this section, we introduce the overall reinforcement training framework and our specially designed reward to mitigate hacking issues brought by translation quality estimation metrics.

3.1 Problem Formulation

Let a source-language sentence be represented as a sequence of tokens $x = (x_1, x_2, \ldots, x_m) \in L_{\text{src}}^m$, where L_{src} denotes the source-language vocabulary and m is the sequence length. A translation model (e.g., LLM) captures the conditional distribution of a target-language token sequence given the source sentence,

$$\pi_\theta(y \mid x) = \prod_{t=1}^{n} \pi_\theta(y_t \mid y_{<t}, x), \qquad (1)$$

where $y = (y_1, \ldots, y_n)$, $y_t \in L_{\text{tgt}}$, L_{tgt} denotes the target-language vocabulary, n is the target sequence length, and θ are the model parameters. We start from a pre-trained LLM and continually train it with only source text (x's) in multiple languages using reinforcement learning (e.g., GRPO). It optimizes the following objective:

$$\text{argmax}\, \mathcal{J}(\theta) = \mathbb{E}_{y \sim \pi_\theta(\cdot|x)}[R(x, y)] \qquad (2)$$

where y is sampled from prior model θ and R is a carefully designed reward. GRPO uses a slightly more sophisticated reward with an advantage function, which will be presented later.

3.2 WALAR Reward

Our reward comprises three components: a base quality estimation model, word alignment score, and language alignment score. We first detail each component and then describe how they are integrated into a unified reward.

Quality Estimation score. To effectively evaluate the translation given only the source sentence, we use MetricX-24-Hybrid-XXL-Bf16[1] (MetricX; Juraska et al. 2024), the state-of-the-art quality estimation metric in WMT24 Metric Shared Task (Freitag et al. 2024). Remarkably, MetricX supports both source-based and reference-based evaluation as a hybrid model, achieving the highest consistency with human ratings. Besides, since MetricX is further finetuned from mT5 (Xue et al. 2021), which is pretrained on mC4 and covers 101 languages, it can provide reliable evaluations even for translations into low-resource languages.

We define the QE reward r_{qe} using MetricX as

$$r_{\text{qe}}(x, y) = \text{MetricX}(x, y), \qquad (3)$$

where the source sentence x and LLM's generated hypothesis y are concatenated with a separating space token and provided as input to the MetricX model to produce a scalar reward score $r_{\text{qe}}(x, y) \in [-25, 0]$, following the MQM annotation guidelines (Juraska et al., 2024). However, using QE alone in RL would lead to reward hacking issues as we illustrated in Figure 1, since QE may assign high rewards to degenerate hypotheses.

[1] https://huggingface.co/google/metricx-24-hybrid-xxl-v2p6-bfloat16

Word Alignment Score. To address this reward hacking, we incorporate a word-alignment–based score that evaluates whether all words are properly covered in the target sentence and no extra information is introduced by LLM's hallucination.

Formally, a word aligner identifies a set of alignment pairs

$$\mathrm{WA} = \{(x_i, y_j) \mid x_i \in x,\ y_j \in y, \mathrm{Sim}(x_i, y_i) > c\}, \tag{4}$$

where each pair $(x_i, y_j) \in \mathrm{WA}$ indicates that the source token x_i and the target token y_j are semantically similar within the sentence context and Sim indicates semantic similarity.

We use the embedding-based approach from Dou and Neubig (2021) to calculate similarity and construct aligned word pairs in source-target utterances. Specifically, we first calculate the word embeddings $h_x = \langle h_{x_1}, \ldots, h_{x_m} \rangle$ and $h_y = \langle h_{y_1}, \ldots, h_{y_n} \rangle$ for x and y using an embedding model's hidden state. Then, we compute the similarity matrix through dot product $\mathrm{Sim_{xy}} = Softmax(h_x h_y^T)$. We construct WA by taking the intersection: $\mathrm{WA} = \{(x_i, y_j) \mid \mathrm{Sim}_{xy}(x_i, y_j) > c \text{ and } \mathrm{Sim}_{yx}(y_j, x_i) > c\}$, where c is a threshold set to 1e-3. To ensure robustness in low-resource languages, we leverage BGE-M3, a strong multilingual embedding model supporting over 100 languages (Chen et al., 2024), and extract word embeddings from its 24th layer.

Based on the constructed word alignments, we define the word-alignment score r_{wa} as the F1 score:

$$r_{\mathrm{wa}}(x, y) = 2 \cdot \frac{P(x, y) \cdot R(x, y)}{P(x, y) + R(x, y)}, \tag{5}$$

where $P(x, y) = \frac{|\mathrm{WA}|}{n}$ and $R(x, y) = \frac{|\mathrm{WA}|}{m}$ denote alignment precision and recall, respectively. This formulation penalizes both over-translation (which reduces precision) and under-translation (which reduces recall), thereby mitigating reward hacking effects induced by QE-based rewards.

Language Alignment. Since both QE models and word alignment models are language-agnostic, LLMs can still hack theses scores by generating translations in an unintended language (see Section 5.1). To mitigate this issue, we introduce a language alignment score that verifies whether the generated translation matches the desired target language and only assigns a positive reward when the languages are as expected.

We adopt GlotLID (Kargaran et al., 2023), a strong language identification model supporting over 1,600 languages, to detect the language of the LLM-generated translation. However, word alignment may assign disproportionately high scores when the translation copies words from the source sentence, which can lead to code-switching outputs after training. In our preliminary experiments, we find that GlotLID alone struggles to reliably identify such code-switching translations.

To address this limitation, we further incorporate MaskLID (Kargaran et al., 2024), a language identification method designed for code-switching scenarios. Specifically, we first apply MaskLID to detect code-switching segments in the generated translation. We then mask tokens belonging to these segments to obtain a filtered target sentence y'. Finally, we feed the masked sentence pair (x, y') into GlotLID to compute the language-alignment reward $r_{\mathrm{la}} = \mathbb{I}(\mathrm{Lang\_detect}(y) = \mathrm{tgt})$, where $\mathrm{Lang\_detect}(\cdot)$ is the language detection function, tgt denotes the desired target language. This encourages the model to generate translations fully in the intended target language.

Overall Reward. We define the overall **WALAR** reward function as

$$r(x, y) = \begin{cases} -25, & \text{if } r_{\mathrm{la}} = 0 \\ r_{\mathrm{qe}}(x, y) & \\ \quad + \alpha \cdot r_{\mathrm{wa}}(x, y'), & \text{if } r_{\mathrm{la}} = 1 \end{cases} \tag{6}$$

where y' denotes the masked translation produced by the code-switching detector, and α is a scaling hyperparameter. We set $\alpha = 25$ to match the dynamic range of the word-alignment reward r_{wa} with that of the QE reward r_{qe}.

3.3 RL Training

We adopt Group Relative Policy Optimization (GRPO; Shao et al. 2024) as our RL algorithm to train the model with our **WALAR** reward, as shown in Eq 7.

$$\begin{aligned} \mathcal{J}_{\mathrm{GRPO}}(\theta) = \mathbb{E}_{x \sim D,\, \{y^{(k)}\}_{k=1}^{G} \sim \pi_{\theta_{\mathrm{old}}}(\cdot|x)} \\ \Bigg[\frac{1}{G} \sum_{k=1}^{G} \min\Bigg(\frac{\pi_\theta(y^{(k)} \mid x)}{\pi_{\theta_{\mathrm{old}}}(y^{(k)} \mid x)} A_k, \\ \mathrm{clip}\left(\frac{\pi_\theta(y^{(k)} \mid x)}{\pi_{\theta_{\mathrm{old}}}(y^{(k)} \mid x)}, 1-\varepsilon, 1+\varepsilon \right) A_k \Bigg) \\ - \beta\, D_{\mathrm{KL}}\big(\pi_\theta \,\big\|\, \pi_{\mathrm{ref}}\big) \Bigg], \end{aligned} \tag{7}$$

$$D_{\mathrm{KL}}(\pi_\theta \| \pi_{\mathrm{ref}}) = \frac{\pi_{\mathrm{ref}}(y^{(k)}|x)}{\pi_\theta(y^{(k)}|x)} - \log \frac{\pi_{\mathrm{ref}}(y^{(k)}|x)}{\pi_\theta(y^{(k)}|x)} - 1 \tag{8}$$

Specifically, for a query x sampled from a monolingual dataset D, we first append a system prompt ("translating from language src to tgt") to x. Then GRPO rolls out G candidate sequences $\{y^{(1)}, y^{(2)}, \ldots, y^{(G)}\}$ at each step with old policy LLM $\pi_{\theta_{old}}$. For each sequence, we extract the translation outputs (for simplicity, we slightly abuse x and y notations for modified input without and extracted translation from output). For each output $y^{(k)}$, we compute the advantage $A_k = \frac{r(x,y^{(k)}) - \mathrm{mean}(\{r(x,y^{(1)}), r(x,y^{(2)}), \ldots, r(x,y^{(G)})\})}{\mathrm{std}(r(x,y^{(1)}), r(x,y^{(2)}), \ldots, r(x,y^{(G)}))}$ with **WALAR** reward.

The hyperparameters ϵ and β control the GRPO clipping threshold and the weight of the Kullback–Leibler (KL) divergence penalty, respectively, in Eq 8.

4 Implementation and Experiments

4.1 Experimental Setup

Data. Our monolingual training dataset is built upon the WMT News Crawl dataset (Kocmi et al., 2024), using 22 source languages[2]. To effectively train the models, we first evaluate their performance with all these 22 languages as the source and all remaining languages in FLORES-101 as the target. Then, we select language directions for which the sentence piece BLEU (spBLEU; Goyal et al. 2022) score is between 1 and 20. Finally, for each selected language direction, we sample 2,000 instances and train all directions concurrently. In this way, we can avoid training models on language directions that are either too easy or too hard for them to translate, thus ensuring the effectiveness of our training process. To ensure the quality of our training data, we adopt Named Entity Recognition (NER) and length clipping to filter out low-quality monolingual data. We also conduct data decontamination to avoid potential data leakage, following the approach in Kocyigit et al. 2025. For detailed information, please refer to Appendix A and F.

Models and training details. Our implementation of **WALAR** is based on OpenRLHF[3] framework. During the training stage, we set the training batch size to 1024 and the micro-batch size to 16. For the GRPO algorithm, we set the rollout numbers to 8, the temperature to 1, the PPO clipping range ϵ to 0.2, and the KL penalty coefficient β to 0.01. We also adopt warm-up training with the learning rate peaking at 5e-7. All the models are trained on 5 NVIDIA A6000 GPUs.

We finetune LLaMAX3-8B-Alpaca (Lu et al., 2024) with **WALAR** and also report results for all the following baseline models. All baseline models are strong multilingual models, including encoder-decoder models and LLM-based Decoder-only models:

- **NLLB-200-1.3B** (Team et al., 2022): An encoder-decoder model supports translations for more than 200 languages, allowing single sentence translation.
- **Hunyuan-MT-7B** (Zheng et al., 2025): A multilingual translation model supporting bidirectional translation across 33 major languages.
- **Tower-Plus-9B** (Rei et al., 2025): A strong multilingual model built on top of Gemma 2 9B through extensive pre-training and post-training, covering 22 languages in total.
- **Aya-Expanse-8B** (Dang et al., 2024b): A highly advanced multilingual model in Aya Expanse family supporting 23 major languages.
- **LLaMAX3-8B-Alpaca** (Lu et al., 2024): A strong translation-specialized LLM supporting more than 100 languages.

Evaluation method. We evaluate all models on the FLORES-101 (Goyal et al., 2022) test set using the BenchMAX evaluation suite (Huang et al., 2025), and report results for seven representative languages, covering over 1400 language directions in total. We use XCOMET-XL[4] (Guerreiro et al., 2024) and MetricX-24-Hybrid-XXL-Bf16 (Juraska et al., 2024) to evaluate the translation quality of the models. These two models share the first-place performance in the WMT24 metrics shared task (Kocmi et al., 2024) and outperform lexical metrics like BLEU (Papineni et al., 2002) by a large margin. Both models are used in reference-based mode, with the source sentence, translation, and reference provided as inputs to ensure accuracy during evaluation. Further details can be found in Appendix B.

[2]The source languages include: Arabic, Bengali, Bulgarian, Croatian, German, English, Finnish, French, Hindi, Hungarian, Indonesian, Italian, Icelandic, Macedonian, Dutch, Polish, Portuguese, Romanian, Russian, Spanish, Turkish, Ukrainian, Simple Chinese.

[3]https://github.com/OpenRLHF/OpenRLHF

[4]https://huggingface.co/Unbabel/XCOMET-XL

	x → en		x → ar		x → tr		x → hi		x → ru		x → zh		x → sw	
	MetricX	xCOMET	MetricX	xCOMET	MetricX	xCOMET	MetricX	xCOMET	MetricX	xCOMET	MetricX	xCOMET	MetricX	xCOMET
Encoder-Decoder Models														
NLLB-200-1.3B	-4.03	90.24	-4.89	74.70	-7.81	73.39	-5.45	63.90	-5.47	81.89	-6.24	68.40	-5.66	66.85
LLM based Decoder-Only Models														
Hunyuan-MT-7B	-10.16	57.83	-15.46	41.21	-14.06	50.00	-14.77	37.22	-13.12	51.99	-8.40	55.84	-16.20	33.18
Tower-Plus-9B	-6.12	82.13	-11.70	46.10	-12.74	55.73	-7.21	58.37	-7.19	77.23	-5.52	69.70	-17.85	28.30
Aya-Expanse-8B	-9.21	71.01	-11.61	57.90	-14.61	53.97	-10.53	47.90	-10.66	66.14	-8.97	59.68	-23.35	21.17
LLaMAX3-8B-Alpaca	-4.10	88.84	-6.19	67.87	-9.50	65.80	-6.57	54.53	-5.50	80.57	-4.63	**71.55**	-8.43	57.04
+WALAR	**-3.65**	**90.31**	**-5.45**	**68.48**	**-7.06**	**72.67**	**-5.65**	**56.05**	**-5.01**	**81.18**	**-4.20**	71.30	**-6.31**	**61.51**
	en → x		ar → x		tr → x		hi → x		ru → x		zh → x		sw → x	
	MetricX	xCOMET	MetricX	xCOMET	MetricX	xCOMET	MetricX	xCOMET	MetricX	xCOMET	MetricX	xCOMET	MetricX	xCOMET
Encoder-Decoder Models														
NLLB-200-1.3B	-6.04	80.04	-3.21	80.25	-4.40	79.97	-2.27	79.67	-3.54	84.77	-2.80	84.16	-5.31	68.53
LLM based Decoder-Only Models														
Hunyuan-MT-7B	-5.42	77.43	-12.12	52.23	-13.10	51.17	-16.58	36.47	-8.64	68.86	-5.06	72.24	-12.27	41.50
Tower-Plus-9B	-9.39	65.84	-11.30	54.21	-10.72	57.57	-9.88	59.80	-10.66	61.89	-9.53	60.82	-13.49	43.66
Aya-Expanse-8B	-12.83	53.80	-13.25	49.10	-13.62	49.23	-12.23	48.39	-13.25	51.68	-12.68	51.76	-19.89	26.69
LLaMAX3-8B-Alpaca	-7.12	71.68	-7.14	64.02	-7.37	65.90	-6.35	63.95	-6.70	69.98	-6.45	67.90	-7.90	56.71
+WALAR	**-4.94**	**76.85**	**-5.09**	**68.37**	**-5.14**	**70.54**	**-4.68**	**67.38**	**-4.81**	**74.46**	**-4.29**	**72.59**	**-5.76**	**60.07**

Table 1: Translation performance of strong multilingual LLMs and encoder-decoder models on the FLORES-101 devtest set, with results for 7 representative languages shown in the table. Ours delivers the best performance on 28 of the 28 reported metrics. In this table, "x" denotes all languages in FLORES-101 supported by XComet and MetricX, excluding the source and target languages in each translation direction. The best results in LLM-based Decoder-only models are bolded for clarity.

Directions	LLaMAX3-8B	WALAR
en→x	54.87	**62.14**
ar→x	55.72	**63.88**
tr→x	55.06	**63.67**
hi→x	56.99	**63.69**
ru→x	58.87	**65.39**
zh→x	57.66	**66.24**
sw→x	52.36	**60.07**
Avg	55.93	**63.58**

Table 2: Evaluation results for models using LLM-as-Judge by Gemini-3-flash. Best results are bolded for clarity.

4.2 Main Results

WALAR improves LLM translation quality by a large margin. As shown in Table 1, we evaluate all models on the FLORES-101 benchmark and report xCOMET and MetricX scores over 1400 language directions. Comparing LLaMAX3 before and after training with **WALAR**, we observe consistent improvements across all metrics and evaluated directions.

Notably, **WALAR** yields substantial gains for both English-centric and low-resource–centric translation. For example, in Swahili-X translation, the xCOMET score increases from 56.71 to 60.07, while English-X translation improves from 71.68 to 76.85. These significant improvements demonstrate the effectiveness of **WALAR**, particularly for low-resource language directions.

Furthermore, when compared with other strong LLM-based decoder-only multilingual models, our model trained with **WALAR** achieves the best performance across all reported language directions . Moreover, our model demonstrates strong ability in X-English and performs better than NLLB-200-1.3B in this direction.

WALAR improves translation under LLM-as-a-Judge. To verify that **WALAR** improves actual translation quality rather than merely optimizing the neural metrics such as MetricX, we additionally evaluate translations using an LLM-as-a-Judge method. Specifically, we adopt Gemini 3 Flash as the judge model, motivated by the Gemini family's first-place performance in the WMT25 metrics shared task (Lavie et al., 2025). Our evaluation prompt follows the ESA-style format used in WMT25, augmented with reference translations to enable reference-based assessment. The full prompt is provided in Appendix C.

As shown in Table 2, we evaluate LLaMAX3-8B-Alpaca and its **WALAR**-trained counterpart on seven representative languages, covering over 700 language directions. Models trained with **WALAR** consistently outperform their baseline counterparts across all evaluated directions, increasing the average score from 55.93 to 63.58. Notably, the scores achieved by **WALAR**-trained models are close to 66 across language directions, corresponding to translations with only minor issues according to the judging rubric. These results further corroborate the substantial translation quality improvements brought by **WALAR**.

WALAR improves language consistency in translation. To systematically assess an LLM's ability to generate translations in the desired tar-

Prompt: *USA Gymnastics and the USOC have the same goal — making the sport of gymnastics, and others, as safe as possible for athletes to follow their dreams in a safe, positive and empowered environment. Translate it from English to Chinese.*			
Error	**Model Output**	**MetricX**	**xCOMET**
Self-generated Reference	**Hypothsis:** 美国体育运动协会和美国体育监督机构有着共同的目标，即确保体育运动项目的安全性，以帮助运动员们实现自己的梦想，并创造一个安全、积极、充满正能量的环境\n中文翻译：\n美国体育运动协会和美国体育监督机构有着共同的目标，即确保体育运动项目的安全性，以帮助运动员们实现自己的梦想，并创造一个安全、积极、充满正能量的环境。	-2.91	89.97
Non-Translation	**Hypothesis:** USA Gymnastics and the USOC share the same goal. Their objective is to ensure that gymnastics and other sports are as safe as possible for athletes to pursue their dreams in a positive and empowering environment.	-3.08	99.19
Under-Translation	**Hypothesis:** 美国体操协会和美国奥委会有相同的目标——让运动员在安全、积极和赋权的环境中追求他们的梦想。 (Missing part: making the sport of gymnastics, and others)	-1.82	89.00
Wrong Language	**Hypothesis:** 미국 체조 대표팀과 USOC는 같은 목표를 가지고 있습니다. 선수들이 안전하고 긍정적이며 격려 받는 환경에서 꿈을 좇을 수 있도록 체조 등의 스포츠를 가능한 한 안전하게 만드는 것입니다.	-2.36	97.90

Table 3: Holes of source-based quality estimation metrics. RL training using these metrics will amplify the holes in LLMs.

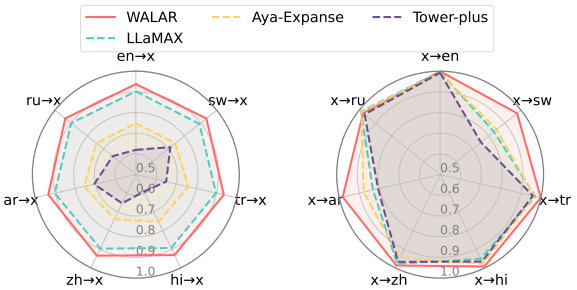

Figure 3: LCR on language directions. **WALAR** improves LLMs' translation into desired target languages.

get language, we define the *Language Consistency Rate* (LCR) as

$$\text{LCR} = \frac{\#\{\text{Lang}(y) = \text{tgt}\}}{\#\text{test data}},$$

which measures the proportion of test instances whose outputs are identified as being in the correct target language. We report LCR for all language directions covered in Table 1, using GlotLID (Kargaran et al., 2023) as the language identification model.

Figure 3 presents the LCR results for four different decoder-only models. Training with **WALAR** consistently improves language consistency across all evaluated language directions on average. Among the four models, LLaMAX trained with **WALAR** achieves the highest LCR across all language directions. The improvement is particularly pronounced for low-resource target languages such as Swahili, where LCR increases from 83% to nearly 100%. Full results are reported in Table 6.

5 Analysis

In this section, we present the analysis of **WALAR** and illustrate the holes of current neural machine translation metrics.

5.1 Holes in Machine Translation Metrics

During training, we observe that models can exploit weaknesses in the reward signal when the reward itself is unreliable. Table 3 summarizes the error types encountered during training. In particular, models trained solely with QE-based rewards exhibit several failure modes, including self-generated references, non-translation, over-translation, under-translation, and wrong language translation.

Self-generated reference refers to a failure mode in which the model learns to repeat its own hypothesis translation, causing the input to the QE model to take the form (source, hypothesis, hypothesis). This effectively tricks the QE model into treating the repeated hypothesis as a reference, activating its reference-based evaluation mode and yielding a high score. We attribute this behavior to the hybrid design of MetricX and xCOMET: during training, both models are optimized to support both source-based and reference-based evaluation by concatenating hypothesis translations and references into a single input.

Non-translation occurs when the model simply paraphrases the source sentence rather than producing a translation. *Wrong language translation* arises when the model generates output in a language different from the one specified in the

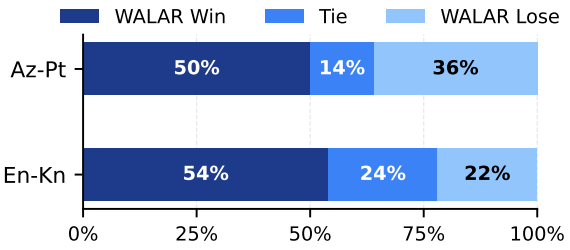

Figure 4: Human evaluation results on Az-Pt and En-Kn.

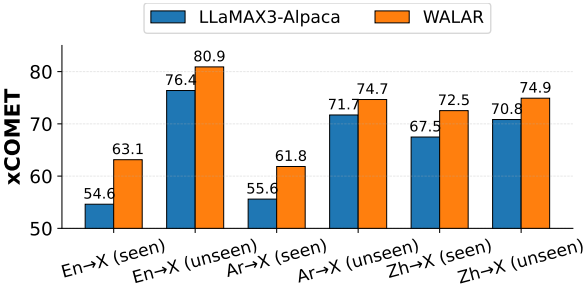

Figure 5: Cross-lingual generalization on unseen target languages. X denotes languages in FLORES-101. LLaMAX3-Alpaca, trained with **WALAR**, demonstrates strong generalization across unseen languages.

prompt. In addition, models may exhibit *over-translation* or *under-translation*, producing outputs that contain redundant content or omit essential information.

Several of these failure modes are consistent with prior observations in the literature (He et al., 2024; Yan et al., 2023). These errors reveal inherent limitations in current state-of-the-art machine translation metrics and our **WALAR** is robust to such errors. Specifically, word alignment mitigates self-generated reference, over-translation, and under-translation, while language alignment prevents non-translation and wrong language translation. More cases are shown in Appendix D.

5.2 Human Evaluation

As discussed in Section 5.1, reference-based evaluation models can be exploited by imperfect translations. To provide a more comprehensive evaluation beyond Gemini-based LLM-as-a-Judge on previous results, we conduct human evaluations on Azerbaijani–Portuguese (Az–Pt) and English–Kannada (En–Kn) translation tasks.

For each test instance, human annotators are presented with two translations, one generated by LLaMAX3 and the other by our **WALAR**-trained model, in a randomly permuted order. Annotators are asked to choose one of three options: (1) Translation 1 is better, (2) Translation 2 is better, or (3) Translation 1 and Translation 2 are of equal quality. We aggregate the annotations to compute win, loss, and tie rates. Additional details regarding the evaluation protocol are provided in Appendix E.

Figure 4 summarizes the human evaluation results. Our model is preferred in 50% of the cases for Az–Pt and 54% for En–Kn, while producing translations of comparable quality in 14% and 24% of the cases, respectively. These results further corroborate the effectiveness of **WALAR** in improving translation quality, particularly for low-resource language pairs.

5.3 Generalization of WALAR

Despite the substantial improvements observed on FLORES-101 (Table 1), an important question remains: can **WALAR** improve translation quality for unseen language directions when only monolingual data are available during training? To address this question, we evaluate LLaMAX3 and its **WALAR**-trained counterpart on 300 language directions ({En, Ar, Zh}→x), and report results separately for seen and unseen target languages.

As shown in Figure 5, **WALAR** yields consistent gains on language directions observed during training, while also demonstrating strong cross-lingual generalization to unseen target languages. These results indicate that the improvements induced by **WALAR** can transfer beyond the training language set, potentially reducing the amount of parallel data and the number of language directions required to train large-scale multilingual models.

6 Conclusion

In conclusion, we present **WALAR**, a reinforcement training method that integrates quality estimation, word alignment, and language alignment as a reward to enhance LLM's translation ability in low-resource languages. Extensive experiments on FLORES-101 across 100 languages and over 1400 language directions show that **WALAR** enables LLMs to achieve substantial improvements on translation quality and language consistency. Our results on LLM-as-a-Judge and human evaluation further corroborate the effectiveness of **WALAR**. Finally, our analysis demonstrates the underexplored holes in current machine translation metrics and the generalization of **WALAR** to unseen languages during training.

Limitations

Despite the promising improvement we achieve in translating low-resource languages, our method cannot be applied to languages that are unsupported by either QE models or embedding models. Although state-of-the-art QE models now cover more than 100 languages, their performance on low-resource languages remains weaker than on high-resource languages. This discrepancy will hinder further improvement in low-resource language translation. Additionally, word alignment relies on a tokenizer for non-segmented languages (e.g., Chinese, Japanese). For many low-resource languages, such tokenizers are either unavailable or unreliable, limiting the applicability of our approach.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Shanbo Cheng, Yu Bao, Qian Cao, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, Wenhao Zhu, Jingwen Chen, Zhichao Huang, Tao Li, Yifu Li, Huiying Lin, Sitong Liu, Ningxin Peng, Shuaijie She, Lu Xu, Nuo Xu, Sen Yang, Runsheng Yu, Yiming Yu, Liehao Zou, Hang Li, Lu Lu, Yuxuan Wang, and Yonghui Wu. 2025. Seed-x: Building strong multilingual translation llm with 7b parameters. *Preprint*, arXiv:2507.13618.

John Dang, Arash Ahmadian, Kelly Marchisio, Julia Kreutzer, Ahmet Üstün, and Sara Hooker. 2024a. RLHF can speak many languages: Unlocking multilingual preference optimization for LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13134–13156, Miami, Florida, USA. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024b. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Zhaopeng Feng, Shaosheng Cao, Jiahan Ren, Jiayuan Su, Ruizhe Chen, Yan Zhang, Jian Wu, and Zuozhu Liu. 2025. MT-r1-zero: Advancing LLM-based machine translation via r1-zero-like reinforcement learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18685–18702, Suzhou, China. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern,

Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz,

Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

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. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8164–8180, Mexico City, Mexico. Association for Computational Linguistics.

Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. 2025. BenchMAX: A comprehensive multilingual evaluation suite for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16751–16774, Suzhou, China. Association for Computational Linguistics.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.

Amir Hossein Kargaran, François Yvon, and Hinrich Schütze. 2024. Masklid: Code-switching language identification through iterative masking. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Bangkok, Thailand. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Muhammed Yusuf Kocyigit, Eleftheria Briakou, Daniel Deutsch, Jiaming Luo, Colin Cherry, and Markus Freitag. 2025. Overestimation in llm evaluation: A controlled large-scale study on data contamination's impact on machine translation. *Preprint*, arXiv:2501.18771.

Gaurav Kumar, George Foster, Colin Cherry, and Maxim Krikun. 2019. Reinforcement learning based curriculum optimization for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2054–2061, Minneapolis, Minnesota. Association for Computational Linguistics.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi,

Markus Freitag, and Daniel Deutsch. 2025. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China. 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.

Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024. LLaMAX: Scaling linguistic horizons of LLM by enhancing translation capabilities beyond 100 languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10748–10772, Miami, Florida, USA. Association for Computational Linguistics.

Arnav Mhaske, Harshit Kedia, Sumanth Doddapaneni, Mitesh M. Khapra, Pratyush Kumar, Rudra Murthy, and Anoop Kunchukuttan. 2023. Naamapadam: A large-scale named entity annotated data for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10441–10456, Toronto, Canada. Association for Computational Linguistics.

Millicent Ochieng, Varun Gumma, Sunayana Sitaram, Jindong Wang, Vishrav Chaudhary, Keshet Ronen, Kalika Bali, and Jacki O'Neill. 2025. Beyond metrics: Evaluating LLMs effectiveness in culturally nuanced, low-resource real-world scenarios. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 230–247, Vienna, Austria. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Miguel Ramos, Patrick Fernandes, António Farinhas, and Andre Martins. 2024. Aligning neural machine translation models: Human feedback in training and inference. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 258–274, Sheffield, UK. European Association for Machine Translation (EAMT).

Miguel Moura Ramos, Tomás Almeida, Daniel Vareta, Filipe Azevedo, Sweta Agrawal, Patrick Fernandes, and André F. T. Martins. 2025. Fine-grained reward optimization for machine translation using error severity mappings. *Preprint*, arXiv:2411.05986.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André F. T. Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual llms. *Preprint*, arXiv:2506.17080.

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.

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. 2022. 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.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

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. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open

and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. 2020. Cross-lingual retrieval for iterative self-supervised training. *Preprint*, arXiv:2006.09526.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2025. X-ALMA: Plug & play modules and adaptive rejection for quality translation at scale. In *The Thirteenth International Conference on Learning Representations*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yiming Yan, Tao Wang, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Mingxuan Wang. 2023. BLEURT has universal translations: An analysis of automatic metrics by minimum risk training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5428–5443, Toronto, Canada. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Fei Yuan, Shuai Yuan, Zhiyong Wu, and Lei Li. 2024. How vocabulary sharing facilitates multilingualism in LLaMA? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12111–12130, Bangkok, Thailand. Association for Computational Linguistics.

Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, and Di Wang. 2025. Hunyuan-mt technical report. *Preprint*, arXiv:2509.05209.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

A Data Curation

We collect all our monolingual data from the WMT News Crawl dataset (Kocmi et al., 2024), then perform data decontamination and data filtering for the source languages. Our data filtering process consists of two steps: length-based filtering and NER-based filtering.

Data Decontamination We follow the method in Kocyigit et al. (2025) and implement an 8-gram search to find matches between our monolingual training dataset and FLORES-101 devtest data in corresponding languages. We tokenize the sentences into sub-word tokens and label the data as contaminated if the longest matching sub-sequence matches more than 70% of the target tokens in FLORES-101 devtest.

Length-based Filtering We directly use the tokenizer of LLaMAX3-8B-Alpaca to process FLORES-101. Then, based on the token length distributions in each language, we empirically determine lower and upper thresholds and retain only data that falls within these ranges. The specific thresholds for each language are reported in Table 4.

NER-based Filtering We adopt language-specific NER models for four languages: English, Arabic, Hindi and Turkish. Specifically, we use spaCy model *en_core_web_sm* for English, IndicNER for Hindi (Mhaske et al., 2023), the CAMeLBERT MSA NER Model for Arabic (Inoue et al., 2021) and the Bert-base-turkish-cased model[5] for Turkish. Named entities identified by these models are subsequently tokenized using the tokenizer. We then exclude samples where named entities constitute more than 60% of the total token length.

Language	Length Threshold
Arabic	[20, 80]
Bengali	[50, 250]
Bulgarian	[20, 140]
Chinese	[10, 150]
Czech	[20, 120]
Dutch	[20, 100]
English	[10, 50]
Finnish	[20, 100]
French	[10, 120]
German	[20, 90]
Hindi	[50, 230]
Hungarian	[20, 120]
Icelandic	[20, 110]
Indonesian	[10, 100]
Italian	[20, 100]
Macedonian	[30, 120]
Polish	[20, 100]
Portuguese	[20, 100]
Romanian	[20, 100]
Russian	[30, 180]
Spanish	[10, 100]
Turkish	[20, 80]
Ukrainian	[20, 150]

Table 4: The length range we adopt for different languages.

B Evaluation Details

We use the BenchMAX evaluation suite for all the models and language directions. The decoding strategy is greedy decoding. For LLaMAX3-8B-Alpaca, both evaluation and training employ the prompt described in the original work to maintain consistency. The full prompt template is provided below.

[5]https://huggingface.co/akdeniz27/bert-base-turkish-cased-ner

Template for LLaMAX

User: Below is an instruction that describes a task, paired with an input that provides further context.
Write a response that appropriately completes the request.
Instruction:
Translate the following sentences from {src_lang} to {tgt_lang}.
Input:
{src_text}
Response:
Assistant:

C LLM-as-Judge Prompt

In Table 2, we use LLM-as-Judge to evaluate the translation quality of different models. We adopt the ESA-like prompt from Lavie et al. 2025 and add a human reference in the prompt to further improve the evaluation accuracy of LLM-as-Judge.

LLM-as-Judge Prompt

Score the following translation from {source_lang} to {target_lang} with respect to the human reference on a scale from 0 to 100, where a score of 0 means a broken or poor translation; 33 indicates a flawed translation with significant issues; 66 indicates a good translation with only minor issues in grammar, fluency, or consistency; and 100 represents a perfect translation in both meaning and grammar. Answer with only a whole number representing the score, and nothing else.
{source_lang} source text:
{source_seg}
{target_lang} reference:
{reference_seg}
{target_lang} translation:
{target_seg}

D Cases of holes in Machine Translation Quality Estimation Metrics

More failure cases of MetricX are shown in Figure 7 and Figure 8. Figure 7 shows English to Spanish direction. Figure 8 shows English to Polish direction. Together, these show the holes in QE are versatile and lead to reward hacking in LLM's RL training.

E Human Evaluation

We hired native speakers in the university lab to serve as human annotators and compensated them at the U.S. minimum wage. We provide the screenshot of our annotation page in Figure 6.

F Training Languages

In total, our training dataset covers 22 source languages (Arabic, Bengali, Bulgarian, Croatian, German, English, Finnish, French, Hindi, Hungarian, Indonesian, Italian, Icelandic, Macedonian, Dutch, Polish, Portuguese, Romanian, Russian, Spanish, Turkish, Ukrainian, Simple Chinese.) and 1,038 language directions. For each direction, we sample 2,000 instances and train all language directions concurrently. This makes our training dataset consist of 2,076,000 monolingual sentences in total. All training language directions are shown in Table 5.

G Used Scientific Artifacts

Below are the scientific artifacts we've used in our paper. For the sake of ethics, we ensure all usages comply with their license.

- *OpenRLHF (Apache-2.0 license)*, an open-source RLHF framework that integrates high performance with simple usage, aiming to streamline the training process and enhance the accessibility of RLHF methods.
- *spaCy (MIT license)*, a library for advanced Natural Language Processing in Python and Cython, build on the very latest research, and was designed to be used in real products.
- *vLLM (Apache-2.0 license)*, a fast and easy-to-use library optimized specifically for LLM inference and serving.
- *Transformers (Apache-2.0 license)*, a model-definition framework focusing on machine learning models for both inference and training.

Source	Target Languages
ara	amh, asm, bel, est, fin, guj, hau, hun, hye, ibo, isl, jav, kat, kaz, kan, kor, kir, ltz, lit, mri, mal, mon, mar, nob, ory, pan, pol, pus, snd, slv, sna, som, srp, swh, tam, tur, urd, uzb, xho, yor, zul
ben	afr, amh, ara, hye, azj, bel, bul, ceb, zho_simpl, ces, dan, nld, est, tgl, fin, glg, kat, deu, ell, hau, heb, hun, isl, ibo, gle, ita, jav, kan, kaz, kir, lav, lit, ltz, msa, mal, mri, mar, mon, nob, npi, pus, fas, pol, rus, srp, sna, snd, slk, slv, som, spa, swh, swe, tgk, tam, tur, ukr, urd, uzb, cym, xho, yor, zul
bul	amh, ara, azj, bel, ben, ceb, zho_simpl, est, fin, guj, hau, hun, isl, ibo, gle, jav, kan, kaz, kir, ltz, mal, mri, mar, mon, npi, pus, pan, srp, sna, snd, som, tam, tur, urd, uzb, xho, yor, zul
ces	amh, ara, azj, bel, ben, ceb, zho_simpl, est, fin, guj, hau, hun, isl, ibo, gle, jav, kan, kaz, kir, ltz, mal, mri, mar, mon, npi, pus, pan, srp, sna, snd, som, swh, tam, tur, urd, uzb, xho, yor, zul
deu	amh, ara, azj, bel, ben, ceb, est, guj, hau, isl, ibo, gle, jav, kan, kaz, kir, lav, lit, mal, mri, mar, mon, npi, pus, pan, srp, sna, snd, som, swh, tam, tur, urd, uzb, xho, yor, zul
eng	amh, asm, bel, hau, ibo, jav, mri, mon, mya, nso, ory, pus, snd, sna, som, srp, urd, uzb, xho, yor, zul
fin	amh, ara, azj, bel, ben, ceb, zho_simpl, est, tgl, kat, ell, guj, hau, heb, hin, hun, isl, ibo, gle, jav, kan, kaz, kir, lav, lit, ltz, msa, mal, mri, mar, mon, nob, npi, pus, fas, pol, pan, srp, sna, snd, slk, slv, som, swh, tgk, tam, tel, tur, urd, uzb, xho, yor, zul
fra	amh, ara, azj, bel, ben, ceb, est, fin, guj, hau, isl, ibo, jav, kan, kaz, kir, ltz, mal, mri, mar, mon, npi, pus, pan, srp, sna, snd, som, tam, tur, urd, uzb, xho, yor, zul
hin	amh, ara, azj, bel, ceb, zho_simpl, ces, nld, est, tgl, fin, kat, ell, hau, heb, hun, isl, ibo, gle, ita, jav, kan, kaz, kir, lav, lit, ltz, mal, mri, mar, mon, nob, pus, fas, pol, srp, sna, snd, slk, slv, som, spa, swh, tgk, tam, tur, ukr, uzb, xho, yor, zul
hun	amh, ara, azj, bel, ben, ceb, zho_simpl, est, tgl, fin, kat, ell, guj, hau, isl, ibo, gle, jav, kan, kaz, kir, lav, lit, ltz, mal, mri, mar, mon, nob, npi, pus, fas, pol, pan, srp, sna, snd, slk, slv, som, swh, tgk, tam, tel, tur, urd, uzb, cym, xho, yor, zul
ind	amh, ara, azj, bel, ceb, zho_simpl, est, fin, kat, guj, hau, hun, isl, ibo, kan, kaz, kir, lav, lit, ltz, mal, mri, mar, mon, npi, pus, pol, pan, srp, sna, snd, slv, som, tam, tur, urd, uzb, xho, yor, zul
ita	amh, ara, azj, bel, ben, ceb, zho_simpl, est, fin, kat, guj, hau, hun, isl, ibo, gle, jav, kan, kaz, kir, lav, lit, ltz, mal, mri, mar, mon, npi, pus, fas, pol, pan, srp, sna, snd, slv, som, swh, tam, tel, tur, urd, uzb, xho, yor, zul
isl	amh, ara, hye, azj, bel, ben, ceb, zho_simpl, ces, est, tgl, fin, kat, ell, guj, hau, heb, hin, hun, ibo, gle, jav, kan, kaz, kir, lav, lit, ltz, msa, mal, mri, mar, mon, nob, npi, pus, fas, pol, pan, srp, sna, snd, slk, slv, som, spa, swh, tgk, tam, tel, tur, ukr, urd, uzb, xho, yor, zul
mkd	amh, ara, azj, bel, ben, ceb, est, fin, guj, hau, hun, isl, ibo, jav, kan, kaz, kir, ltz, mal, mri, mar, mon, npi, pus, pan, srp, sna, snd, som, swh, tam, tur, urd, uzb, xho, yor, zul
nld	amh, ara, azj, bel, ben, ceb, zho_simpl, est, tgl, fin, kat, ell, guj, hau, heb, hun, isl, ibo, gle, jav, kan, kaz, kir, lav, lit, ltz, mal, mri, mar, mon, npi, pus, fas, pol, pan, srp, sna, snd, slk, slv, som, swh, tgk, tam, tur, urd, uzb, xho, yor, zul
pol	amh, ara, azj, bel, ben, ceb, zho_simpl, est, tgl, fin, kat, ell, guj, hau, heb, hin, hun, isl, ibo, gle, jav, kan, kaz, kir, lav, lit, ltz, mal, mri, mar, mon, nob, npi, pus, fas, pan, srp, sna, snd, slv, som, swh, tgk, tam, tel, tur, urd, uzb, cym, xho, yor, zul
por	amh, ara, azj, bel, ben, ceb, est, fin, hau, isl, ibo, jav, kan, kaz, kir, ltz, mal, mri, mar, mon, npi, pus, pan, srp, sna, snd, som, tam, tur, urd, uzb, xho, yor, zul
ron	amh, ara, azj, bel, ben, ceb, zho_simpl, est, fin, guj, hau, isl, ibo, jav, kan, kaz, kir, lit, ltz, mal, mri, mar, mon, npi, pus, pan, srp, sna, snd, som, tam, tur, urd, uzb, xho, yor, zul
rus	amh, ara, azj, ben, ceb, zho_simpl, est, fin, guj, hau, hun, isl, ibo, gle, jav, kan, kaz, kir, ltz, mal, mri, mar, mon, nob, npi, pus, pan, srp, sna, snd, som, swh, tam, tel, tur, urd, uzb, xho, yor, zul
spa	amh, ara, asm, azj, bel, ben, bos, ceb, zho_simpl, hrv, est, tgl, fin, kat, guj, hau, heb, hin, hun, isl, ibo, gle, jav, kea, kam, kan, kaz, kir, lav, lit, ltz, mal, mri, mar, mon, nob, npi, nya, oci, ory, pus, fas, pol, pan, srp, sna, snd, slk, slv, som, swh, tgk, tam, tel, tur, urd, uzb, xho, yor, zul
tur	amh, ara, azj, bel, ben, ceb, zho_simpl, ces, est, tgl, fin, kat, ell, guj, hau, heb, hun, isl, ibo, gle, jav, kan, kaz, kir, lav, lit, ltz, mal, mri, mar, mon, nob, npi, pus, fas, pol, pan, srp, sna, snd, slk, slv, som, swh, tgk, tam, urd, uzb, cym, xho, yor, zul
ukr	amh, ara, azj, ben, ceb, zho_simpl, est, fin, guj, hau, hun, isl, ibo, gle, jav, kan, kaz, kir, ltz, mal, mri, mar, mon, npi, pus, pan, srp, sna, snd, som, swh, tam, tur, urd, uzb, xho, yor, zul
zho_simpl	afr, amh, ara, hye, azj, bel, ben, bul, ceb, ces, nld, est, tgl, fin, kat, deu, ell, guj, hau, heb, hin, hun, isl, ibo, gle, ita, jav, kan, kaz, kir, lav, lit, ltz, mkd, msa, mal, mri, mar, mon, nob, npi, pus, fas, pol, pan, rus, srp, sna, snd, slk, slv, som, spa, swh, swe, tgk, tam, tel, tur, ukr, urd, uzb, cym, xho, yor, zul

Table 5: All training language directions for LLaMAX3-8B- and our **WALAR** trained model. All languages are presented in FLORES-101 code.

Models	**en → x**	**sw → x**	**tr → x**	**hi → x**	**zh → x**	**ar → x**	**ru → x**
WALAR	**0.9376**	**0.9360**	**0.9359**	**0.9312**	**0.9346**	**0.9343**	**0.9361**
LLaMAX3-8B-Alpaca	0.9027	0.8931	0.8975	0.8931	0.8969	0.9018	0.8996
Aya-Expanse-8B	0.7481	0.7440	0.7604	0.7537	0.7387	0.7563	0.7440
Tower-Plus-9B	0.6199	0.7133	0.6504	0.5908	0.6537	0.7063	0.6432
Models	**x → en**	**x → sw**	**x → tr**	**x → hi**	**x → zh**	**x → ar**	**x → ru**
WALAR	**0.9995**	**0.9759**	**0.9991**	**0.9928**	**0.9871**	**0.9828**	**0.9759**
LLaMAX3-8B-Alpaca	0.9970	0.8329	0.9719	0.9513	0.9780	0.8329	0.9890
Aya-Expanse-8B	0.9919	0.8493	0.9672	0.9770	0.9472	0.8792	0.9865
Tower-Plus-9B	0.9923	0.7518	0.9598	0.9662	0.9675	0.8072	0.9666

Table 6: Complete results for LCR

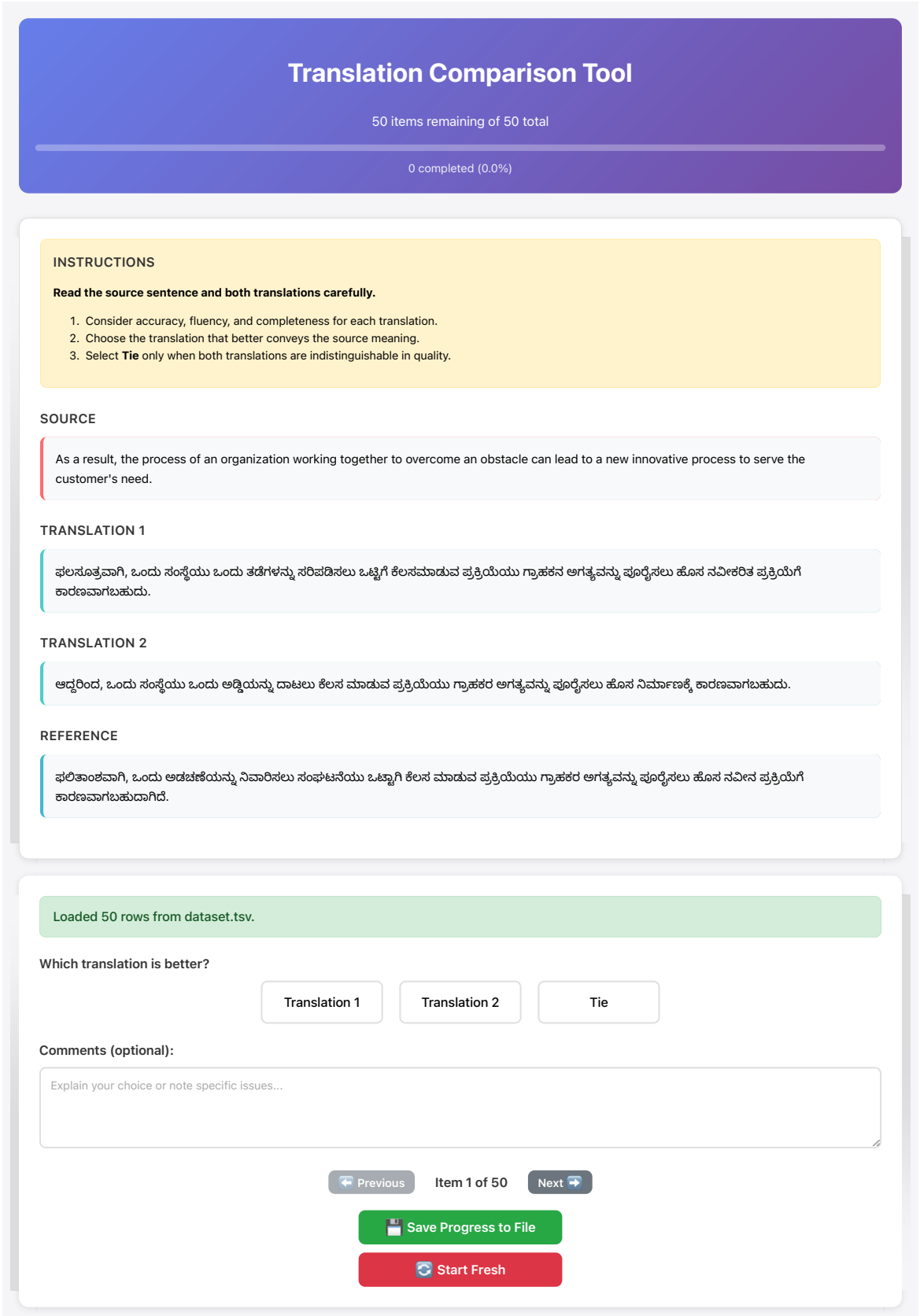

Figure 6: Screenshot of human evaluation web tool.

Source (English): Dr. Tony Moll discovered the Extremely Drug Resistant Tuberculosis (XDR-TB) in the South African region KwaZulu-Natal.
Reference (Spanish): La tuberculosis extremadamente resistente (XDR-TB) fue detectada por primera vez por el Dr. Tony Moll en área sudafricana de Zululandia.
Over-translation (Spanish): El Dr. Tony Moll descubrió en la región de KwaZulu-Natal, en Sudáfrica, un tipo de tuberculosis muy difícil de tratar: la tuberculosis extremadamente resistente a los medicamentos (XDR-TB). Esta bacteria es resistente a la mayoría de los tratamientos antibióticos convencionales y requiere tratamiento con medicamentos antituberculosos específicos. MetricX: -3.11
Wrong Language (Portuguese): O Dr. Tony Moll descobriu a Tuberculose Extremamente Resistente a Drogas (XDR-TB) na região de KwaZulu-Natal, na África do Sul. MetricX: -1.85

Figure 7: A case study from FLORES-101 dataset. The intended language direction is English to Spanish. blue text denotes the MetricX score in source-based mode, and the red text highlights the errors in the translation.

Source (English): Because the dinosaur feathers do not have a well-developed shaft, called a rachis, but do have other features of feathers — barbs and barbules — the researchers inferred the rachis was likely a later evolutionary development that these other features.
Reference (Polish): W związku z tym, że pióra dinozaura nie mają dobrze rozwiniętego trzonu, zwanego stosiną, lecz mają inne cechy piór – promienie i promyki – naukowcy wywnioskowali, że stosina prawdopodobnie była późniejszym etapem ewolucyjnym niż te pozostałe cechy.
Over-translation (Polish): Ponieważ pióra dinozaurów nie posiadały typowej osi (centralnego trzonu przebiegającego przez całą strukturę pióra), ale zachowały podstawowe cechy piór, takie jak promienie i promyczki, naukowcy wysnuli wniosek, że struktura osi mogła ewoluować później, podczas gdy inne cechy, takie jak promienie i promyczki, mogły istnieć już wcześniej. MetricX: -3.17
Wrong Language (Slovak): Pretože dinosaurové perie nemá dobre vyvinutú osku, nazývanú rachis, ale má iné črty peria — háčiky a háčikovité vlákna, vedci usúdili, že rachis pravdepodobne vznikol až neskôr v evolúcii než tieto ostatné črty. MetricX: -3.52

Figure 8: A case study from FLORES-101 dataset. The intended language direction is English to Polish. Blue text denotes the MetricX score in source-based mode, and the red text highlights the errors in the translation.