

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

Yifeng Liu¹, Siqi Ouyang¹, Yatish Hosmane Revanasiddappa¹, Lei Li¹

¹ Carnegie Mellon University

{yifengl, siqiouya, yhosmane}@andrew.cmu.edu, leili@cs.cmu.edu

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 LLMs 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 1,414 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

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 from English to Chinese.

Output 1: 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。中文翻译：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。
(Self-generated Reference, xCOMET:89.97, MetricX: -2.91)

Output 2: 미국 체조 대표팀과 USOC는 같은 목표를 가지고 있습니다. 선수들이 안전하고 긍정적이며 격려 받는 환경에서 꿈을 쫓을 수 있도록 체조 등의 스포츠를 가능한 한 안전하게 만드는 것입니다. (Wrong Language, xCOMET: 97.90, MetricX: -2.36)

Output 3: 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. (Non-translation, xCOMET: 99.19, MetricX: -3.08)

Output 4: 美国体操协会和美国奥委会的目标是确保运动员在一个安全、积极且充满支持的环境中追求他们的体操梦想。他们的共同目标是使体操这项运动以及其他运动项目尽可能安全，让运动员能够专注于训练和比赛，而不必担心安全问题。
(Over-translation, xCOMET: 81.06, MetricX: -2.05)

Figure 1: A case from English to Chinese showing the holes of quality estimation metrics. RL training using these metrics will amplify the holes in LLMs.

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.

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 to use 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 — comparing 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 (Fritag et al., 2024), these QEs exhibit noticeable holes when applied to LLM training, such as failure to detect over- and under-translation, wrong language words and self-generated reference. Figure 1 illustrates examples of xCOMET and 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 in the LLM just repeating input source sentences. Astonishingly, a 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 the group relative policy optimization (GRPO) training framework and post-train LLMs based on Qwen3-8B (Team, 2025), LLaMAX3-8B-Alpaca (Lu et al., 2024) and Translategemma-4B-it (Finkelstein et al., 2026). 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 three LLMs using our **WALAR**. Our experiments demonstrate that our models outperform the strongest prior LLM of the same size in 1,414 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. Closely related to this line of work, Pombal et al. (2025) systematically studies metric interference, showing that reusing the same or related automatic metrics for quality-guided decoding can severely distort instance-level metric scores and reduce their agreement with human judgments.

Multilingual LLMs Recent progress in LLMs has continuously increased the number of languages they support (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, fall short in low-resource languages due to the scarcity of

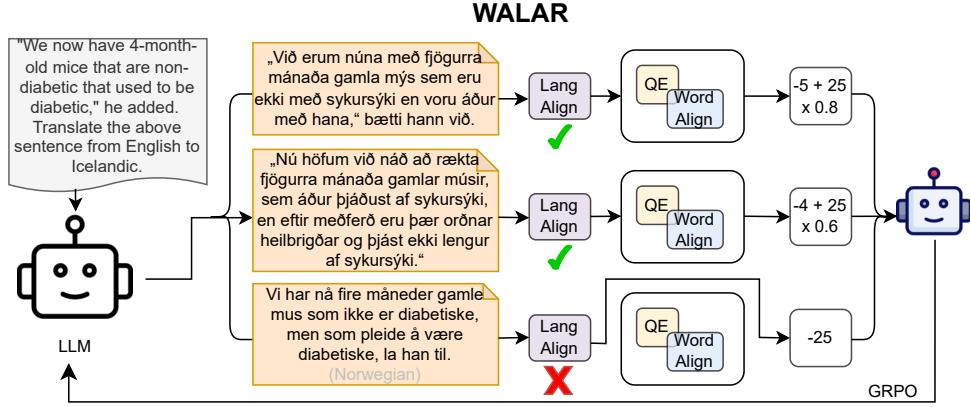


Figure 2: Illustration of **WALAR**. On each step, the LLM is prompted to translate one monolingual sentence into another language with several different rollouts. Each output will then be evaluated by language alignment, quality estimation, and word alignment. Finally, the LLM is trained using GRPO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

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.

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 $x = (x_1, x_2, \dots, x_m) \in L_{\text{src}}^m$, 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,

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

where $y = (y_1, \dots, y_n)$, $y_t \in L_{\text{tgt}}$, 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:

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

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-Bf16¹ (MetricX; Juraska et al. 2024), the state-of-the-art quality estimation metric in WMT24 Metric Shared Task (Fritag 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

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

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

$$\text{WA} = \{(x_i, y_j) \mid x_i \in x, y_j \in y, \text{Sim}(x_i, y_j) > c\}, \quad (4)$$

where each pair $(x_i, y_j) \in \text{WA}$ 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 $h_x = \langle h_{x_1}, \dots, h_{x_m} \rangle$ and $h_y = \langle h_{y_1}, \dots, h_{y_n} \rangle$ for x and y using an embedding model’s hidden state. Then, we compute the similarity matrix through dot product $\text{Sim}_{xy} = \text{Softmax}(h_x h_y^T)$. We construct WA by taking the intersection: $\text{WA} = \{(x_i, y_j) \mid \text{Sim}_{xy}(x_i, y_j) > c \text{ and } \text{Sim}_{yx}(y_j, x_i) > c\}$, 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:

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

where $P(x, y) = \frac{|\text{WA}|}{n}$ and $R(x, y) = \frac{|\text{WA}|}{m}$ 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 these 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 y' into GlotLID to compute the language-alignment reward $r_{\text{la}} = \mathbb{I}(\text{Lang_detect}(y') = \text{tgt})$, where $\text{Lang_detect}(\cdot)$ 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

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

where y' denotes the masked translation produced by the code-switching detector, and α is a scaling hyperparameter set to 20.

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.

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

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

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 $\{y^{(1)}, y^{(2)}, \dots, y^{(G)}\}$ at each step with old policy LLM $\pi_{\theta_{\text{old}}}$. 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 $A_k = \frac{r(x, y^{(k)}) - \text{mean}(\{r(x, y^{(1)}), r(x, y^{(2)}), \dots, r(x, y^{(G)})\})}{\text{std}(r(x, y^{(1)}), r(x, y^{(2)}), \dots, r(x, y^{(G)}))}$ 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 23 source languages². To effectively train the models, we first evaluate their performance with these 23 languages as the source and all other FLORES-101 languages supported by MetricX as the target. Then, we select language directions for which the sentence piece BLEU (sp-BLEU; Goyal et al. 2022) score is between 1 and 20. Finally, for each selected language direction, we sample 250 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 I.

²The 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, Simplified Chinese.

Models and training details. Our implementation of **WALAR** is based on OpenRLHF³ 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 report results for strong multilingual encoder-decoder models and LLM-based decoder-only models. For the encoder-decoder model, we include NLLB-200-1.3B (Team et al., 2022). For LLM-based decoder-only models, we evaluate Hunyuan-MT-7B (Zheng et al., 2025), Tower-Plus-9B (Rei et al., 2025), Aya-Expanse-8B (Dang et al., 2024b), Qwen3-8B in non-thinking mode (Team, 2025), Translatagemma-4B-it (Finkelstein et al., 2026) and LLaMAX3-8B-Alpaca (Lu et al., 2024), among which we further finetune LLaMAX3-8B-Alpaca, Qwen3-8B in non-thinking mode and Translatagemma-4B-it with **WALAR**. Moreover, we employ another strong baseline LLaMAX3-8B-Alpaca+**WALAR-SFT**, which is a supervised finetuned model trained with high-scoring translations selected by **WALAR**’s reward as pseudo-references. Specifically, we sample 32 possible translations for each sentence with min_p=0.01 and select the translation with the highest **WALAR**’s reward as the pseudo-reference. Then, we finetune LLaMAX3-8B-Alpaca with the pseudo-references using cross-entropy loss.

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 central languages⁴, covering 1,414 language directions in total. These languages are selected to represent their respective language families across diverse geographic and demographic contexts. We use sp-BLEU (Goyal et al., 2022), XCOMET-XL⁵ (Guerreiro et al., 2024), MetricX-24-Hybrid-XXL-Bf16 (Juraska et al., 2024) and Gemini 3 Flash (Team et al., 2025) in reference-based mode to evaluate the translation quality of the models. To prevent LLMs from exploiting the neural metrics by gen-

³<https://github.com/OpenRLHF/OpenRLHF>

⁴English, Arabic, Turkish, Hindi, Russian, Simplified Chinese, Swahili.

⁵<https://huggingface.co/Unbabel/XCOMET-XL>

spBLEU	x → en	en → x	x → ar	ar → x	x → tr	tr → x	x → hi	hi → x	x → ru	ru → x	x → zh	zh → x	x → sw	sw → x	Avg
NLLB-200-1.3B Δ	39.03	30.23	24.91	22.57	23.30	22.47	24.51	22.22	25.80	22.18	18.71	18.40	24.37	21.77	24.32
HunYuan-MT-7B	21.04	14.32	16.29	9.84	15.37	10.72	13.19	9.66	16.17	10.46	15.55	9.83	8.18	7.46	12.72
Tower-Plus-9B	31.55	15.32	9.42	10.69	13.36	11.15	20.53	11.39	23.06	11.54	23.01	9.98	3.29	9.46	14.55
Aya-Expand-8B	24.03	14.25	14.59	10.33	12.73	10.62	14.07	10.20	17.22	10.72	15.49	9.29	2.36	4.71	12.19
Qwen3-8B	29.22	16.67	15.70	11.11	14.00	11.45	12.10	11.48	19.05	12.08	21.08	10.58	1.29	6.72	13.75
+WALAR	28.59	17.11	15.71	11.89	14.73	12.38	12.03	12.05	19.31	12.68	20.87	11.08	3.35	7.76	14.25
Translatagemma-4B-it	27.45	18.19	18.86	12.58	16.39	13.05	17.85	12.75	20.62	13.74	19.46	11.39	10.17	11.61	16.01
+WALAR	28.26	19.81	19.06	13.70	16.65	14.34	17.76	14.41	21.11	14.75	19.90	12.21	12.44	13.24	16.97
LLaMAX3-8B-Alpaca	32.24	21.34	17.50	14.14	12.23	14.89	16.64	15.42	21.48	16.20	18.37	12.99	13.21	15.12	17.27
+WALAR-SFT	32.78	22.72	17.95	16.51	16.82	16.49	18.15	16.60	22.10	17.39	19.13	14.52	16.04	16.30	18.82
+WALAR	32.56	23.68	17.81	17.14	18.02	17.86	18.00	17.61	22.45	17.90	20.08	15.23	17.15	17.35	19.49
xCOMET*	x → en	en → x	x → ar	ar → x	x → tr	tr → x	x → hi	hi → x	x → ru	ru → x	x → zh	zh → x	x → sw	sw → x	Avg
NLLB-200-1.3B Δ	90.08	79.00	62.78	70.50	73.12	72.77	62.91	69.44	81.46	76.23	65.36	70.49	61.29	61.93	71.24
HunYuan-MT-7B	76.65	44.47	52.51	39.38	67.48	40.22	55.90	38.29	72.59	41.92	67.69	41.72	36.56	27.32	50.19
Tower-Plus-9B	82.03	45.13	37.16	43.72	55.09	42.99	57.83	40.90	76.78	45.00	68.53	44.91	20.36	35.95	49.74
Aya-Expand-8B	70.89	45.04	48.73	42.01	53.54	42.02	47.52	41.50	66.00	44.36	58.36	43.35	18.83	22.45	46.04
Qwen3-8B	83.88	57.85	52.64	51.18	60.66	52.08	43.63	51.38	73.77	56.58	70.45	54.91	16.45	31.31	54.06
+WALAR	84.77	62.98	55.15	56.07	64.55	57.13	48.27	55.78	76.71	61.17	71.37	59.61	21.84	35.17	57.90
Translatagemma-4B-it	87.10	65.40	56.63	57.41	71.01	58.77	60.93	56.22	81.25	62.73	72.57	60.97	44.62	48.64	63.16
+WALAR	87.73	70.04	57.39	61.72	72.25	63.11	60.25	62.38	81.48	67.26	72.60	65.38	51.43	53.00	66.14
LLaMAX3-8B-Alpaca	88.67	68.66	57.61	61.38	65.10	63.10	53.50	61.34	80.51	66.93	70.29	64.80	53.68	54.00	64.97
+WALAR-SFT	89.16	71.48	59.98	64.19	68.43	66.13	55.84	63.93	81.11	69.99	71.09	67.87	56.09	56.45	67.27
+WALAR	90.44	76.42	68.06	68.48	73.36	71.05	58.56	68.03	82.39	74.65	71.99	72.80	62.20	60.31	71.34
MetricX*	x → en	en → x	x → ar	ar → x	x → tr	tr → x	x → hi	hi → x	x → ru	ru → x	x → zh	zh → x	x → sw	sw → x	Avg
NLLB-200-1.3B Δ	-3.46	-4.48	-4.57	-4.07	-7.57	-4.87	-5.55	-4.71	-3.70	-4.22	-8.06	-4.97	-3.89	-4.28	-4.89
HunYuan-MT-7B	-7.80	-14.75	-11.57	-15.03	-10.05	-15.21	-8.42	-15.03	-8.30	-14.97	-5.78	-14.75	-14.85	-17.23	-12.41
Tower-Plus-9B	-6.07	-14.85	-14.21	-14.32	-12.74	-14.74	-7.33	-14.48	-7.18	-14.59	-5.75	-14.09	-19.81	-15.78	-12.14
Aya-Expand-8B	-9.14	-15.00	-13.89	-15.10	-14.49	-15.32	-10.56	-14.23	-10.60	-14.98	-9.12	-14.93	-23.50	-20.66	-14.68
Qwen3-8B	-6.12	-11.82	-11.17	-12.35	-11.57	-12.50	-9.54	-11.21	-8.29	-11.78	-5.60	-11.60	-22.86	-17.54	-10.99
+WALAR	-5.79	-10.03	-10.49	-10.29	-10.45	-10.55	-8.97	-9.41	-7.42	-9.97	-5.23	-9.67	-21.01	-15.71	-9.93
Translatagemma-4B-it	-5.53	-8.90	-5.99	-9.89	-8.65	-9.57	-6.25	-9.12	-5.60	-9.60	-4.95	-10.55	-10.38	-10.77	-8.27
+WALAR	-5.12	-6.70	-5.78	-7.01	-7.87	-7.34	-5.98	-6.33	-5.16	-7.11	-4.42	-8.44	-8.47	-8.17	-6.71
LLaMAX3-8B-Alpaca	-4.08	-7.69	-9.08	-7.81	-9.62	-8.03	-6.90	-7.08	-5.49	-7.45	-4.97	-7.24	-9.37	-8.62	-7.53
+WALAR-SFT	-3.99	-6.87	-8.44	-6.85	-8.53	-7.04	-6.29	-6.25	-5.30	-6.55	-4.75	-6.22	-8.54	-7.71	-6.67
+WALAR	-3.60	-5.35	-6.01	-5.44	-7.02	-5.46	-5.66	-4.98	-4.77	-5.15	-4.27	-4.71	-6.76	-6.10	-5.38
Gemini*	x → en	en → x	x → ar	ar → x	x → tr	tr → x	x → hi	hi → x	x → ru	ru → x	x → zh	zh → x	x → sw	sw → x	Avg
LLaMAX3-8B-Alpaca	74.61	52.86	51.14	53.71	57.18	52.99	61.04	54.84	66.08	56.51	64.18	55.43	50.80	50.14	57.25
+WALAR	78.75	63.45	63.25	64.83	68.68	65.22	70.18	65.55	70.00	67.60	69.93	67.28	62.51	61.17	67.03

Table 1: Model performance on FLORES-101 test set, with results for 7 central languages shown in the table. Δ denotes encoder-decoder models. Bold text denotes the best result across LLM-based decoder-only models. For spBLEU and Gemini*, we evaluate on all 102 languages covered in FLORES-101. For xCOMET* and MetricX*, we evaluate only on the languages they support in FLORES-101, covering 81 and 86 languages, respectively. The full language coverage is shown in Appendix J.

erating wrong language translations (Section 5.1), we adopt GlotLID to identify the language of each translation candidate. Candidates identified as being in the wrong language are penalized by assigning the minimum score of the neural metric. We denote this penalized variant of xCOMET, MetricX and Gemini-based LLM-as-a-Judge as xCOMET*, MetricX* and Gemini*, respectively. We evaluate xCOMET* and MetricX* only on languages they support, and spBLEU and Gemini* on all languages in FLORES-101. We also conduct human evaluation to further strengthen our results (Section 5.4). Further details can be found in Appendix B.

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 spBLEU, xCOMET* and MetricX* scores over 1,414 language directions. Comparing Qwen3-8B,

Translatagemma-4B-it and LLaMAX3-8B-Alpaca before and after training with **WALAR**, we observe significant average improvements across all metrics, demonstrating the generalizability of **WALAR** across different model families.

Notably, **WALAR** yields substantial gains for both English-centric and low-resource-centric translation. For example, within the LLaMAX family, **WALAR** improves the xCOMET* score for Swahili-X from 54.00 to 60.31, and for English-X translation from 68.66 to 76.42. These significant improvements demonstrate the effectiveness of **WALAR**, particularly for low-resource language directions. We additionally provide the qualitative examples in Appendix G, Borda count (Himmi et al., 2024) in Appendix D, and results of representative high- and low-resource languages in Appendix E.

WALAR improves translation under LLM-as-a-Judge. To verify that **WALAR** improves actual translation quality rather than merely optimizing

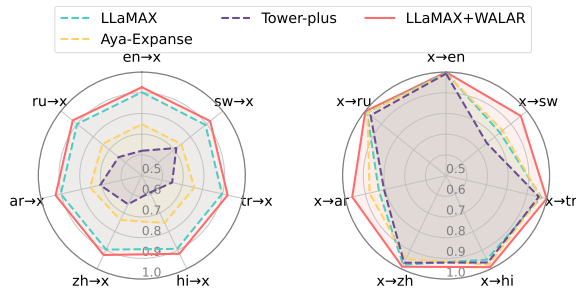


Figure 3: LCR on language directions. **WALAR** improves LLMs’ translation into desired target languages.

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 1, LLaMAX3-8B-Alpaca trained with **WALAR** consistently outperforms its baseline counterpart across all evaluated directions, increasing the average score from 57.25 to 67.03. Notably, the average score achieved by **WALAR**-trained LLaMAX3-8B-Alpaca is higher than 66, 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 target language, we define the *Language Consistency Rate* (LCR) as

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

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, LLaMAX3-8B-

Alpaca 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 9.

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. Figure 1 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. Several of these failure modes are consistent with prior observations in the literature (He et al., 2024; Yan et al., 2023).

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

We also provide the statistical analysis of each error category in Table 2, covering all language directions evaluated in Table 1. Specifically, self-generated reference happens in our preliminary experiments on Qwen2.5-0.5B-Instruct with QE

	Reference	LLaMAX	+QE	+QE+Lang Align	+WALAR
Language Consistency Rate (LCR)	100.00	92.33	7.57	96.04	95.56
Translation Token Length	62.04	83.02	40.00	79.09	65.24

Table 2: Statistics for different types of errors.

Model	Lang→x		x→Lang	
	spBLEU	xCOMET*	spBLEU	xCOMET*
LLaMAX3-8B-Alpaca	15.73	65.73	18.81	69.46
+QE (w/ filter)	1.47	0.01	3.27	0.13
+QE+Lang Align (w/ filter)	12.38	70.84	13.71	72.95
+WALAR (w/o filter)	17.75	70.09	20.73	71.62
+WALAR (w/ filter)	18.11	70.25	20.87	72.43

Table 3: Ablation on the reward components of **WALAR** and spBLEU-based data filtering. **Lang** denotes seven representative languages (English, Arabic, Turkish, Hindi, Russian, Simplified Chinese, Swahili).

only reward. The model exhibits such behavior in 100% of cases. But it does not happen on larger base models like Qwen3-8B and LLaMAX3-8B-Alpaca. For wrong language translation, we measure LCR for all four models with different reward configurations. Results are shown in Table 2. The model trained with QE-only reward exhibits an LCR of 7.57 as the QE model lacks the ability to tell whether the translation is in the right language direction. Language alignment score effectively fixes this issue. For over- and under-translation, we measure the average token length of the generated translations. **WALAR** is the only method whose translation length closely matches the reference, whereas other methods exhibit noticeable length deviation. This confirms that incorporating word alignment into the reward is critical to prevent both omission and over-generation.

5.2 Ablation Study

We conduct the ablation study to demonstrate the contribution of each component in **WALAR**. As shown in Table 3, we train the LLaMAX3-8B-Alpaca with three different rewards: (1) Quality estimation score, (2) Quality estimation score and language alignment, and (3) Quality estimation score, word alignment score and language alignment (**WALAR**). We also use **WALAR** to train the LLaMAX3-8B-Alpaca with data not filtered by spBLEU heuristics (Section 4.1). All models are trained with the same settings described in Section 4.1 and evaluated on the same 1,414 language directions in Table 1.

Results show that the LLaMAX trained with

only the quality estimation score performs worst on both spBLEU and xCOMET*, primarily due to wrong language translations. Adding language alignment improves xCOMET* scores but degrades spBLEU, as it tends to over-translate. In contrast, **WALAR** achieves the best performance on both metrics, demonstrating the importance of word alignment score and language alignment. Additionally, LLaMAX3-8B-Alpaca trained on all language directions performs slightly worse than its counterpart trained on spBLEU-filtered language directions, demonstrating the superiority of spBLEU-based data filtering.

5.3 Effects of Word Alignment Score Weight

The hyperparameter α controls the weight of word alignment reward in **WALAR** (Eq 6). In this subsection, we focus on the question: How to select the best α for our model’s training? To answer this question, we train the LLaMAX3-8B-Alpaca with six different α : 0, 5, 10, 15, 20, 25, and evaluate all the checkpoints on FLORES-101 validation set with spBLEU, MetricX*, and xCOMET*. As illustrated in Table 4, by increasing α from 0 to 20, the spBLEU improves steadily from 12.88 to 19.71, while the MetricX* and xCOMET* degrade. $\alpha = 25$ shows the worst performance across all three metrics. Finally, we report results for $\alpha = 20$ as the hyperparameter in our experiments. We prioritize spBLEU for model selection for two reasons. First, spBLEU is more reliable for low-resource languages because it is a rule-based metric that relies on a fixed multilingual tokenizer rather than a learned neural model. Second, xCOMET and MetricX may be susceptible to over-optimization, since our training procedure directly optimizes toward neural metrics, which can lead to metric inflation.

5.4 Human Evaluation

As discussed in Section 5.1, neural metrics 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

	x → Lang (Avg.)			Lang → x (Avg.)			Overall Avg.		
	spBLEU	MetricX*	xCOMET*	spBLEU	MetricX*	xCOMET*	spBLEU	MetricX*	xCOMET*
$\alpha=0$	13.44	-4.44	74.99	12.31	-4.31	72.17	12.88	-4.38	73.58
$\alpha=5$	19.56	-4.94	74.99	17.08	-4.69	72.98	18.32	-4.82	73.99
$\alpha=10$	20.54	-5.45	74.42	17.88	-4.88	72.46	19.21	-5.17	73.44
$\alpha=15$	20.98	-5.32	73.92	18.12	-4.97	71.78	19.55	-5.15	72.85
$\alpha=20$	21.10	-5.44	73.39	18.32	-5.10	71.66	19.71	-5.27	72.53
$\alpha=25$	18.42	-5.53	71.93	14.89	-5.16	70.25	16.66	-5.35	71.09

Table 4: Performance of LLaMAX3-8B-Alpaca trained with different α on FLORES-101 validation set. We select and report the results of $\alpha = 20$ in Table 1.

tasks.

For each test instance, human annotators are presented with two translations, one generated by LLaMAX3-8B-Alpaca 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 are provided in Appendix H.

Figure 4 summarizes the human evaluation results. Our model is preferred in 42% of the cases for Az–Pt and 51% for En–Kn, while producing translations of comparable quality in 34% and 39% of the cases, respectively. These results further corroborate the effectiveness of **WALAR** in improving translation quality, particularly for low-resource language pairs.

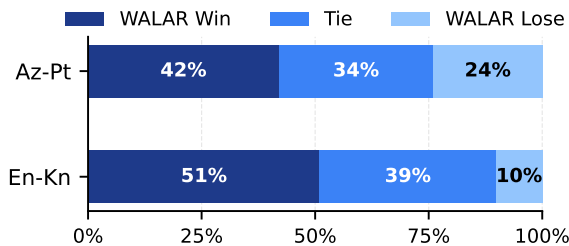


Figure 4: Human evaluation results on Az-Pt and En-Kn.

5.5 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 LLaMAX3-8B-Alpaca and its **WALAR**-trained counterpart on 303 language directions ($\{En, Ar, Zh\} \rightarrow x$), and report

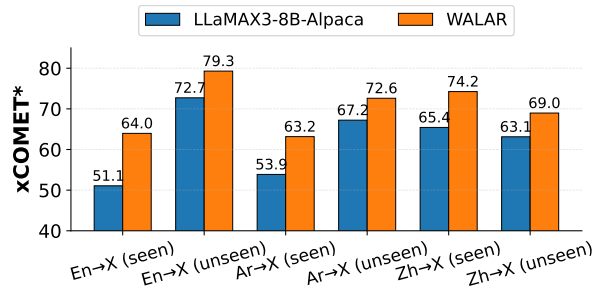


Figure 5: Cross-lingual generalization on unseen target languages. X denotes languages in FLORES-101. LLaMAX3-8B-Alpaca, trained with **WALAR**, demonstrates strong generalization across unseen languages.

results 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 102 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 neural 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 expand: 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, Jiahua 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.
- Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, Markus Freitag, and David Vilar. 2026. [TranslateGemma technical report](#). *Preprint*, arXiv:2601.09012.
- Markus Freitag, Nitika Mathur, Daniel Deutsch, Chikiu 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 Alonso, 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, Jun-teng 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-Yearly, 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 Kam-bador, 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 Rapparthi, 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, Vitor 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, Xuwei 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 Papanikos, 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 Damraj, 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 Khanelwal, 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 Ratanandani, 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.
- Anas Himmi, Ekhine Irurozki, Nathan Noiry, Stephan Cléménçon, and Pierre Colombo. 2024. [Towards more robust NLP system evaluation: Handling missing scores in benchmarks](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11759–11785, Miami, Florida, USA. 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, Nurpeis 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 Sindhuja, 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.
- José Pombal, Nuno M. Guerreiro, Ricardo Rei, and André F. T. Martins. 2025. [Adding chocolate to mint: Mitigating metric interference in machine translation](#). *Preprint*, arXiv:2503.08327.
- 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 Teixeira, 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.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong

Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayanan Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Mon-

teiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturk, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Vilella, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Inuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopoulos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh,

James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiehzadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shunting Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bülle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuwei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrz, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen

Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Praatek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Nicolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzdankowski, Fiona Macintosh, Roopali Vij, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Reynald Chung, Kai Yang, Nihal Balani, Arthur Brażniskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonia Stella, Liselotte Snijders, Michael Mandl, Ante Kärroman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jigeng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Lu

oweï Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnappalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasmurthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebecca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck,

Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Pöder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshv, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurusurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhuy Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshiti Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Sha-

- har Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Avieli Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Píkus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhanian, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. 2025. [Gemini: A family of highly capable multimodal models](#). *Preprint*, arXiv:2312.11805.
- 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.
- Qwen Team. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- 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, Jiayi 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 Kocmygit 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 Qwen3-8B 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 5.

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 model⁶ 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.

B Evaluation Details

We use the BenchMAX evaluation suite for all the models and language directions. The decoding strategy is greedy decoding for LLM-based decoder-only models and beam search for NLLB-200-1.3B (beam size=5, length penalty=0.6). For LLaMAX3-8B-Alpaca, both evaluation and training use the prompt described in the original work to maintain consistency. The full prompt template is provided below.

⁶<https://huggingface.co/akdeniz27/bert-base-turkish-cased-ner>

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 5: The length range we adopt for different languages.

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-a-Judge Prompt

In Table 1, we use LLM-as-a-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-a-Judge.

LLM-as-a-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 Results of Borda Count on FLORES-101

We report the Borda count (Himmi et al., 2024) of each model in Table 8 by averaging the rank across language pairs.

E Results of Representative Languages

To further analyze which languages benefit most and least from WALAR, we select 10 high-resource languages (English, French, German, Italian, Portuguese, Russian, Spanish, Swedish, Simplified Chinese, and Dutch) and 10 low-resource languages (Belarusian, Filipino, Lao, Marathi, Nepali, Sindhi, Somali, Swahili, Punjabi, and Xhosa), based on the amount of available bitext reported in the FLORES-101 paper (Goyal et al., 2022).

F More Cases of Holes in Machine Translation Metrics

More failure cases of MetricX are shown in Figure 8, Figure 9, Figure 10 and Figure 11. Together, these examples show that the holes in QE are versatile and lead to reward hacking during reinforcement training.

G Qualitative Examples of WALAR

We add qualitative translation examples illustrating how WALAR improves LLMs’ translation quality relative to the baselines in Figure 12, Figure 13, Figure 14 and Figure 15, with xCOMET scores provided for reference.

H 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 7.

I Training Languages

In total, our training dataset covers 23 source languages (Arabic, Bengali, Bulgarian, Croatian, German, English, Finnish, French, Hindi, Hungarian, Indonesian, Italian, Icelandic, Macedonian, Dutch, Polish, Portuguese, Romanian, Russian, Spanish, Turkish, Ukrainian, Simplified Chinese). We remove target languages that are either not supported by MetricX or not segmented by spaces (except for Simplified Chinese and Traditional Chinese, for which we use HanLP⁷ to tokenize the sentence). For each direction, we sample 250 instances and train all language directions concurrently.

⁷<https://github.com/hankcs/HanLP>

	x → en spBLEU	en → x spBLEU	x → ar spBLEU	ar → x spBLEU	x → tr spBLEU	tr → x spBLEU	x → hi spBLEU	hi → x spBLEU	x → ru spBLEU	ru → x spBLEU	x → zh spBLEU	zh → x spBLEU	x → sw spBLEU	sw → x spBLEU
Hunyuan-MT-7B	9.317	6.505	4.812	6.604	4.475	6.208	7.762	6.921	8.683	6.604	8.446	5.782	5.762	6.950
Tower-Plus-9B	3.238	6.762	9.238	6.604	7.861	6.495	2.386	6.208	2.901	6.584	2.515	6.594	7.624	6.366
Aya-Expansive-8B	7.020	6.426	6.426	6.089	7.158	6.158	6.436	6.307	6.842	6.228	7.683	6.218	8.782	8.337
Qwen3-8B	4.941	7.020	6.089	7.386	7.119	7.515	8.713	7.455	7.079	7.396	3.475	6.921	9.851	8.079
+WALAR	6.050	6.465	6.822	6.376	6.228	6.386	8.901	6.752	7.257	6.683	4.069	6.168	7.455	6.624
Translatagemma-4b-it	7.515	6.000	2.980	6.248	3.950	6.178	3.525	6.059	5.594	6.059	5.525	6.356	4.891	6.099
+WALAR	6.515	4.287	2.792	4.554	3.475	4.406	3.851	4.257	4.238	4.406	4.752	5.069	3.465	3.970
LLaMAX3-8B-Alpaca	4.129	5.168	5.802	5.257	8.594	5.485	5.881	5.040	5.525	4.901	7.614	5.416	3.861	4.139
+W-SFT	2.762	3.693	4.485	3.307	3.960	3.871	3.455	3.653	3.733	3.317	6.178	3.713	2.099	2.861
+WALAR	3.515	2.673	5.554	2.574	2.178	2.297	4.089	2.347	3.149	2.822	4.743	2.762	1.208	1.574
	x → en xCOMET*	en → x xCOMET*	x → ar xCOMET*	ar → x xCOMET*	x → tr xCOMET*	tr → x xCOMET*	x → hi xCOMET*	hi → x xCOMET*	x → ru xCOMET*	ru → x xCOMET*	x → zh xCOMET*	zh → x xCOMET*	x → sw xCOMET*	sw → x xCOMET*
Hunyuan-MT-7B	6.938	5.700	5.450	5.750	3.825	5.875	4.188	6.125	5.350	5.950	4.325	5.825	6.350	7.912
Tower-Plus-9B	5.275	6.700	9.650	6.987	9.150	6.875	3.812	6.600	4.388	6.662	5.362	6.725	7.950	7.225
Aya-Expansive-8B	7.675	7.050	6.487	6.550	8.200	6.600	7.362	6.725	7.700	7.037	8.488	6.900	8.625	8.838
Qwen3-8B	6.900	7.513	7.862	7.487	8.037	7.662	9.338	7.625	8.700	7.487	5.250	7.412	9.838	7.987
+WALAR	5.225	5.912	5.850	6.025	5.987	6.125	8.137	6.175	6.800	6.100	3.513	5.838	7.225	6.812
Translatagemma-4b-it	5.275	4.487	5.375	4.963	3.587	5.125	2.688	5.125	3.513	4.750	4.062	4.912	5.013	4.537
+WALAR	4.375	3.263	4.588	3.625	2.275	3.663	3.413	3.337	3.163	3.587	3.450	3.587	3.837	3.163
LLaMAX3-8B-Alpaca	6.263	6.100	5.650	5.912	6.575	5.825	6.625	5.862	6.562	5.812	7.763	6.013	3.138	4.112
+W-SFT	5.000	4.812	3.087	4.562	4.775	4.487	5.388	4.425	5.300	4.537	6.487	4.603	2.025	2.850
+WALAR	2.075	3.288	1.000	2.950	2.587	2.625	4.050	2.850	3.525	2.950	5.400	2.975	1.000	1.387
	x → en MetricX*	en → x MetricX*	x → ar MetricX*	ar → x MetricX*	x → tr MetricX*	tr → x MetricX*	x → hi MetricX*	hi → x MetricX*	x → ru MetricX*	ru → x MetricX*	x → zh MetricX*	zh → x MetricX*	x → sw MetricX*	sw → x MetricX*
Hunyuan-MT-7B	6.635	6.365	6.494	6.329	4.318	6.306	5.176	6.741	5.282	6.388	4.412	5.988	6.106	7.882
Tower-Plus-9B	5.247	6.847	9.106	7.153	8.765	7.035	4.518	6.800	4.965	6.812	6.094	6.529	7.035	6.835
Aya-Expansive-8B	8.024	7.141	8.071	7.094	8.435	6.976	8.035	7.071	8.071	7.259	8.765	6.824	9.941	8.812
Qwen3-8B	6.859	7.435	7.647	7.553	7.753	7.576	8.918	7.600	8.224	7.435	5.729	7.012	9.035	7.965
+WALAR	5.306	5.765	6.247	5.824	5.718	6.024	7.835	5.882	6.494	5.918	4.212	5.212	7.718	6.706
Translatagemma-4b-it	5.071	4.141	2.435	4.400	3.953	4.341	3.129	4.529	3.753	4.753	4.259	6.024	4.035	3.882
+WALAR	4.741	2.741	2.482	2.553	2.882	2.882	2.894	2.435	3.176	2.776	3.247	2.612	2.612	2.729
LLaMAX3-8B-Alpaca	5.718	6.024	6.012	5.847	6.365	5.918	6.388	5.871	6.388	5.729	7.329	5.329	4.141	4.306
+W-SFT	4.624	4.694	4.212	4.494	4.388	4.518	4.776	4.506	5.176	4.365	6.212	3.776	2.894	3.188
+WALAR	2.776	3.247	2.294	3.024	2.424	2.824	3.318	3.000	3.459	2.965	4.741	2.376	1.482	1.859

Table 6: Average rank of strong multilingual LLMs on the FLORES-101 test set, with results for 7 central languages shown in the table.

	x → en spBLEU	en → x spBLEU	x → ar spBLEU	ar → x spBLEU	x → tr spBLEU	tr → x spBLEU	x → hi spBLEU	hi → x spBLEU	x → ru spBLEU	ru → x spBLEU	x → zh spBLEU	zh → x spBLEU	x → sw spBLEU	sw → x spBLEU
Hunyuan-MT-7B	34.75	36.50	26.02	25.84	25.13	26.81	19.57	23.68	26.49	27.11	25.14	24.58	12.52	16.76
Tower-Plus-9B	45.14	45.30	13.96	31.17	19.89	32.18	29.02	33.57	34.03	33.93	33.38	28.88	4.90	25.87
Aya-Expansive-8B	42.94	39.08	29.74	30.77	25.81	29.98	23.96	28.28	31.14	30.28	28.62	26.04	3.22	10.47
Qwen3-8B	40.92	37.68	23.12	26.91	21.04	27.10	16.96	25.11	28.02	28.00	29.94	24.58	2.48	15.48
+WALAR	39.94	37.31	22.51	26.48	20.80	26.83	16.70	24.78	27.50	27.62	29.52	23.87	4.44	15.49
Translatagemma-4b-it	36.87	36.93	25.22	26.70	21.96	26.55	22.53	25.84	27.87	28.08	26.26	23.16	14.44	23.96
+WALAR	37.40	38.17	25.30	26.87	22.05	27.32	21.93	26.29	28.51	28.38	26.39	23.13	16.06	24.86
LLaMAX3-8B-Alpaca	39.69	33.48	22.04	24.57	18.03	25.79	20.59	24.62	27.11	27.55	23.03	22.88	18.59	25.20
+W-SFT	39.89	36.15	22.18	27.00	21.51	26.49	21.45	25.20	27.79	28.13	23.85	23.26	20.25	25.82
+WALAR	39.57	36.97	21.62	27.07	21.99	27.66	20.67	26.28	27.56	28.35	24.72	23.62	20.32	26.35
	x → en xCOMET*	en → x xCOMET*	x → ar xCOMET*	ar → x xCOMET*	x → tr xCOMET*	tr → x xCOMET*	x → hi xCOMET*	hi → x xCOMET*	x → ru xCOMET*	ru → x xCOMET*	x → zh xCOMET*	zh → x xCOMET*	x → sw xCOMET*	sw → x xCOMET*
Hunyuan-MT-7B	96.58	95.58	69.21	88.47	89.83	89.14	73.45	84.69	94.19	92.98	87.90	92.25	49.08	62.51
Tower-Plus-9B	97.33	95.31	45.13	87.22	69.22	88.29	70.92	86.56	94.35	93.28	84.73	92.67	24.23	70.25
Aya-Expansive-8B	96.47	91.06	71.01	85.53	80.12	85.66	65.73	82.06	92.38	89.05	81.70	88.37	20.22	36.36
Qwen3-8B	96.65	91.52	64.27	83.96	74.56	84.53	50.84	79.67	90.02	89.58	84.01	89.06	18.10	50.54
+WALAR	96.81	92.73	66.53	85.26	78.10	85.87	55.56	81.21	90.91	90.54	84.89	89.83	22.63	53.47
Translatagemma-4b-it	96.68	94.67	66.50	86.72	83.87	87.35	68.72	83.65	93.09	92.07	83.62	90.89	53.98	75.36
+WALAR	96.62	94.57	66.51	86.77	84.80	87.75	67.52	84.10	92.97	92.09	83.33	91.04	61.02	75.98
LLaMAX3-8B-Alpaca	96.05	89.84	65.06	83.93	74.53	84.52	56.74	80.11	89.65	89.18	78.06	88.07	62.67	75.24
+W-SFT	96.18	90.88	67.39	84.92	77.66	85.63	59.15	81.16	90.14	89.94	78.69	88.90	64.80	76.13
+WALAR	96.73	92.41	77.21	86.20	82.93	87.60	62.44	83.05	91.20	91.17	79.76	90.48	71.43	77.71
	x → en MetricX*	en → x MetricX*	x → ar MetricX*	ar → x MetricX*	x → tr MetricX*	tr → x MetricX*	x → hi MetricX*	hi → x MetricX*	x → ru MetricX*	ru → x MetricX*	x → zh MetricX*	zh → x MetricX*	x → sw MetricX*	sw → x MetricX*
Hunyuan-MT-7B	-1.88	-1.44	-7.33	-2.11	-3.78	-2.53	-3.69	-2.83	-1.72	-1.82	-1.90	-1.27	-11.34	-7.35
Tower-Plus-9B	-1.64	-1.48	-12.04	-2.62	-8.85	-2.92	-3.64	-2.49	-1.78	-1.70	-2.33	-1.19	-18.38	-6.08
Aya-Expansive-8B	-2.00	-2.63	-6.73	-2.87	-6.18	-3.34	-4.29	-3.36	-2.44	-2.69	-2.81	-2.09	-23.34	-15.71
Qwen3-8B	-2.00	-2.72	-8.12	-3.39	-7.51	-3.83	-5.76	-4.02	-3.16	-2.72	-2.34	-2.03	-22.65	-11.55
+WALAR	-1.92	-2.37	-7.56	-3.06	-6.61	-3.47	-5.06	-3.68	-2.91	-2.49	-2.21	-1.83	-20.91	-10.30
Translatagemma-4b-it	-1.82	-0.85	-2.83	-1.39	-5.37	-2.16	-3.55	-2.00	-1.54	-1.45	-2.23	-4.12	-6.60	-2.46
+WALAR	-1.94	-1.04	-3.62	-1.35	-5.50	-2.21	-3.69	-2.00	-1.71	-1.27	-2.04	-3.99	-5.41	-2.36
LLaMAX3-8B-Alpaca	-2.25	-3.17	-7.90	-3.42	-7.73	-3.91	-5.19	-3.99	-3.42	-2.82	-3.51	-2.25	-8.00	-4.65
+W-SFT	-2.19	-2.89	-7.32	-3.18	-6.85	-3.62	-4.74	-3.71	-3.27	-2.60	-3.34	-2.03	-7.38	-4.39
+WALAR	-1.95	-2.38	-4.74	-2.83	-5.46	-3.04	-4.27	-3.30	-2.93	-2.28	-2.99	-1.68	-5.66	-3.84

Table 7: Results of high-resource representative languages, including English, French, German, Italian, Portuguese, Russian, Spanish, Swedish, Simplified Chinese, and Dutch.

J Language Coverage

In our experiments, we evaluate our models with xCOMET* and MetricX* on the languages they support. We list all their supported languages below in Figure 6. Specifically, xCOMET and MetricX cover 81 and 86 languages in FLORES-101, respectively.

K 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.

	x → en spBLEU	en → x spBLEU	x → ar spBLEU	ar → x spBLEU	x → tr spBLEU	tr → x spBLEU	x → hi spBLEU	hi → x spBLEU	x → ru spBLEU	ru → x spBLEU	x → zh spBLEU	zh → x spBLEU	x → sw spBLEU	sw → x spBLEU
Hunyuan-MT-7B	16.22	7.64	12.86	5.22	12.13	5.67	11.90	5.50	12.59	5.47	12.11	5.24	6.42	4.03
Tower-Plus-9B	27.65	4.61	8.20	3.60	11.49	4.14	19.19	4.11	19.56	3.94	19.84	3.62	2.76	4.05
Aya-Expanses-8B	14.64	5.16	9.64	3.34	8.81	3.95	11.49	4.64	10.74	3.67	10.73	3.29	1.91	2.65
Qwen3-8B	26.08	9.08	12.94	4.98	11.88	5.80	11.34	7.86	16.10	6.52	18.48	5.44	0.97	3.90
+WALAR	25.14	10.01	13.95	6.76	13.09	7.38	13.00	8.88	16.44	7.44	18.22	6.50	3.05	5.06
Translatogemma-4b-it	25.98	11.56	17.23	8.02	15.42	8.41	18.70	8.86	19.10	8.90	18.08	7.23	9.01	7.05
+WALAR	26.95	13.12	17.80	9.28	15.73	9.73	18.91	12.74	19.39	10.17	18.62	8.12	11.92	8.83
LLaMAX3-8B-Alpaca	31.44	17.25	17.18	10.52	10.62	11.30	18.27	14.04	20.35	12.43	17.85	9.50	12.31	11.61
+W-SFT	32.07	18.56	17.47	13.14	16.21	13.47	19.62	15.43	20.80	13.98	18.40	11.49	15.85	13.30
+WALAR	31.75	19.31	17.06	14.15	17.32	14.87	19.52	16.21	21.07	14.62	19.05	12.64	17.17	14.62
	x → en xCOMET*	en → x xCOMET*	x → ar xCOMET*	ar → x xCOMET*	x → tr xCOMET*	tr → x xCOMET*	x → hi xCOMET*	hi → x xCOMET*	x → ru xCOMET*	ru → x xCOMET*	x → zh xCOMET*	zh → x xCOMET*	x → sw xCOMET*	sw → x xCOMET*
Hunyuan-MT-7B	59.52	19.51	39.16	15.62	49.60	15.42	44.44	16.15	54.82	17.30	51.24	16.99	27.46	8.85
Tower-Plus-9B	67.68	10.90	30.20	15.02	43.52	14.05	47.53	12.34	60.71	13.33	53.96	14.56	19.43	15.33
Aya-Expanses-8B	48.39	18.99	33.03	18.43	35.89	18.27	35.26	20.82	44.80	19.36	39.93	18.33	17.07	16.54
Qwen3-8B	70.48	32.19	40.54	25.55	47.30	27.24	36.37	32.76	58.73	30.79	56.66	28.38	14.64	17.70
+WALAR	71.87	39.99	44.56	34.50	51.26	35.05	42.88	39.25	62.09	38.12	57.76	36.07	21.07	22.28
Translatogemma-4b-it	78.93	40.81	48.25	34.32	59.65	33.76	55.97	33.50	70.45	36.61	62.72	35.31	38.44	25.61
+WALAR	80.01	47.96	49.55	40.56	60.95	40.68	55.52	48.94	70.86	45.43	62.82	41.27	44.93	33.99
LLaMAX3-8B-Alpaca	80.87	53.47	50.80	45.83	55.52	46.76	51.64	50.76	70.65	50.33	61.65	47.14	46.94	41.01
+W-SFT	82.33	56.27	53.00	48.71	58.81	49.80	53.56	53.33	71.30	53.27	62.59	50.63	49.47	43.36
+WALAR	84.39	62.57	59.75	53.59	63.22	55.48	55.46	57.37	72.81	59.04	63.41	56.51	55.33	47.77
	x → en MetricX*	en → x MetricX*	x → ar MetricX*	ar → x MetricX*	x → tr MetricX*	tr → x MetricX*	x → hi MetricX*	hi → x MetricX*	x → ru MetricX*	ru → x MetricX*	x → zh MetricX*	zh → x MetricX*	x → sw MetricX*	sw → x MetricX*
Hunyuan-MT-7B	-10.38	-19.19	-13.55	-19.36	-12.76	-19.57	-10.38	-19.16	-11.10	-19.26	-7.58	-19.15	-16.61	-21.02
Tower-Plus-9B	-8.17	-22.15	-15.32	-19.95	-14.64	-20.36	-9.17	-20.54	-9.75	-20.82	-7.41	-20.12	-20.33	-19.72
Aya-Expanses-8B	-12.98	-20.03	-16.90	-20.04	-17.87	-20.13	-12.77	-18.17	-15.04	-19.79	-11.70	-20.14	-23.66	-22.25
Qwen3-8B	-8.11	-15.58	-13.05	-17.03	-13.52	-16.44	-12.00	-14.11	-10.61	-15.79	-7.32	-16.21	-22.87	-19.76
+WALAR	-7.63	-12.99	-11.78	-13.52	-12.29	-13.63	-9.91	-11.79	-9.59	-13.21	-6.68	-13.28	-20.79	-17.63
Translatogemma-4b-it	-6.84	-13.11	-7.15	-14.79	-11.08	-13.45	-7.21	-13.07	-7.29	-14.01	-5.83	-15.09	-13.17	-15.22
+WALAR	-6.68	-10.31	-6.88	-10.90	-9.80	-10.39	-6.77	-7.95	-7.16	-10.43	-5.24	-11.87	-11.32	-11.54
LLaMAX3-8B-Alpaca	-4.65	-8.74	-9.35	-9.05	-10.44	-9.22	-6.98	-7.29	-6.25	-8.61	-5.44	-8.73	-9.89	-9.42
+W-SFT	-4.55	-7.85	-8.70	-7.79	-9.21	-7.95	-6.46	-6.39	-6.06	-7.56	-5.19	-7.36	-8.98	-8.34
+WALAR	-4.11	-5.71	-6.38	-5.75	-7.68	-5.71	-5.87	-4.82	-5.44	-5.51	-4.74	-5.11	-7.15	-6.07

Table 8: Results of low-resource representative languages, including Belarusian, Filipino, Lao, Marathi, Nepali, Sindhi, Somali, Swahili, Punjabi, and Xhosa.

- *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.

xCOMET: Afrikaans, Amharic, Arabic, Armenian, Assamese, Azerbaijani, Belarusian, Bengali, Bosnian, Bulgarian, Burmese, Catalan, Chinese, Chinese, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, Galician, German, Greek, Gujarati, Hausa, Hebrew, Hindi, Hungarian, Icelandic, Indonesian, Irish, Italian, Japanese, Javanese, Georgian, Kazakh, Khmer, Kannada, Korean, Kurdish, Kyrgyz, Lao, Latvian, Lithuanian, Macedonian, Malay, Malayalam, Marathi, Mongolian, Nepali, Norwegian, Oriya, Oromo, Pashto, Persian, Polish, Portuguese, Punjabi, Romanian, Russian, Serbian, Sindhi, Slovak, Slovenian, Somali, Spanish, Swahili, Swedish, Tamil, Telugu, Thai, Filipino, Turkish, Ukrainian, Urdu, Uzbek, Vietnamese, Welsh, Xhosa

MetricX: Afrikaans, Danish, Dutch, German, Icelandic, Luxembourgish, Norwegian, Swedish, English, Catalan, French, Galician, Portuguese, Romanian, Spanish, Belarusian, Bulgarian, Czech, Macedonian, Polish, Russian, Serbian, Slovak, Slovenian, Ukrainian, Bengali, Gujarati, Hindi, Marathi, Nepali, Punjabi, Sindhi, Urdu, Armenian, Greek, Irish, Welsh, Italian, Latvian, Lithuanian, Pashto, Kurdish, Tajik, Cebuano, Filipino, Indonesian, Javanese, Malay, Maori, Igbo, Northern Sotho, Shona, Swahili, Umbundu, Xhosa, Yoruba, Zulu, Amharic, Arabic, Maltese, Somali, Azerbaijani, Kazakh, Kyrgyz, Turkish, Uzbek, Kannada, Malayalam, Tamil, Telugu, Burmese, Estonian, Finnish, Hungarian, Georgian, Hausa, Hebrew, Japanese, Khmer, Vietnamese, Korean, Lao, Thai, Mongolian, Traditional Chinese, Simplified Chinese

Figure 6: The supported languages of xCOMET and MetricX.

Translation Comparison Tool

50 items remaining of 50 total

0 completed (0.0%)

INSTRUCTIONS

Read the source sentence and both translations carefully.

1. Consider accuracy, fluency, and completeness for each translation.
2. Choose the translation that better conveys the source meaning.
3. Select **Tie** only when both translations are indistinguishable in quality.

SOURCE

As a result, the process of an organization working together to overcome an obstacle can lead to a new innovative process to serve the customer's need.

TRANSLATION 1

ಫಲಸೂತ್ರವಾಗಿ, ಒಂದು ಸಂಸ್ಥೆಯು ಒಂದು ತಡೆಗಳನ್ನು ಸರಿಪಡಿಸಲು ಒಟ್ಟಿಗೆ ಕೆಲಸಮಾಡುವ ಪ್ರಕ್ರಿಯೆಯು ಗ್ರಾಹಕನ ಅಗತ್ಯವನ್ನು ಪೂರೈಸಲು ಹೊಸ ನವೀಕರಿತ ಪ್ರಕ್ರಿಯೆಗೆ ಕಾರಣವಾಗಬಹುದು.

TRANSLATION 2

ಆದ್ದರಿಂದ, ಒಂದು ಸಂಸ್ಥೆಯು ಒಂದು ಅಡ್ಡಿಯನ್ನು ದಾಟಲು ಕೆಲಸ ಮಾಡುವ ಪ್ರಕ್ರಿಯೆಯು ಗ್ರಾಹಕರ ಅಗತ್ಯವನ್ನು ಪೂರೈಸಲು ಹೊಸ ನಿರ್ಮಾಣಕ್ಕೆ ಕಾರಣವಾಗಬಹುದು.

REFERENCE

ಫಲಿತಾಂಶವಾಗಿ, ಒಂದು ಅಡಚಣೆಯನ್ನು ನಿವಾರಿಸಲು ಸಂಘಟನೆಯು ಒಟ್ಟಿಗೆ ಕೆಲಸ ಮಾಡುವ ಪ್ರಕ್ರಿಯೆಯು ಗ್ರಾಹಕರ ಅಗತ್ಯವನ್ನು ಪೂರೈಸಲು ಹೊಸ ನವೀನ ಪ್ರಕ್ರಿಯೆಗೆ ಕಾರಣವಾಗಬಹುದಾಗಿದೆ.

Loaded 50 rows from dataset.tsv.

Which translation is better?

Translation 1

Translation 2

Tie

Comments (optional):

Explain your choice or note specific issues...

Previous

Item 1 of 50

Next

Save Progress to File

Start Fresh

Figure 7: 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 8: A case study from FLORES-101 dataset. The intended language direction is from 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 promyczki – 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 háčikovitú vlákna, vedci usúdili, že rachis pravdepodobne vznikol až neskôr v evolúcii než tieto ostatné črty. MetricX: -3.52

Figure 9: 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.

Source (English): Previously, Ring’s CEO, Jamie Siminoff, remarked the company started when his doorbell wasn’t audible from his shop in his garage. Translate it from English to Chinese.
Reference (Chinese): 此前，铃声 (Ring) 公司 CEO 杰米·西米诺夫曾表示，公司成立之初，他在位于车库的工作室里是听不到门铃响声的。
Over-translation (Chinese): 以前，Ring公司的CEO Jamie Siminoff提到，公司最初成立时， 其主要业务是通过门铃来吸引顾客。 他的店铺位于车库内，而门铃则安装在车库的入口处。 MetricX: -3.70
Wrong Language (Japanese): 以前、RingのCEOであるジェイミー・シミノフ氏は、自宅のガレージにある作業場からドアベルの音が聞こえなかったことがきっかけで同社を創業したと述べました。 MetricX: -3.59
Non-translation (English): Earlier, Jamie Siminoff, CEO of Ring, shared the story behind the company’s founding; it all began when he noticed that the doorbell at his house wasn’t loud enough for him to hear it clearly from his workshop, which was located in his garage. MetricX: -0.66

Figure 10: A case study from FLORES-101 dataset. The intended language direction is from English to Chinese. **Blue** text denotes the MetricX score in source-based mode, and the **red** text highlights the errors in the translation.

Source (French): Les travailleurs doivent souvent obtenir l’approbation de leurs supérieurs pour toute décision qu’ils prennent et sont censés obéir aux instructions de ces derniers sans les remettre en question.
Reference (German): Oftmals müssen Arbeitnehmer für alle Entscheidungen, die sie treffen, die Zustimmung ihrer Vorgesetzten einholen und es wird von ihnen erwartet, dass sie den Anweisungen ihrer Chefs bedingungslos Folge leisten.
Over-translation (German): Die Mitarbeiter müssen vor jeder Entscheidung die Zustimmung ihrer Vorgesetzten einholen und sich strikt an deren Anweisungen halten, ohne diese willkürlich zu ändern oder deren Ausführung zu verweigern. MetricX: -0.29
Wrong Language (Italian): I lavoratori devono spesso ottenere l’approvazione dei loro superiori per qualsiasi decisione prendano e sono tenuti a seguire le loro istruzioni senza metterle in discussione. MetricX: -0.35
Non-translation (English): Les employés sont généralement tenus de solliciter l’accord de leurs responsables pour chaque décision et doivent suivre leurs directives sans les contester. MetricX: -0.37

Figure 11: A case study from FLORES-101 dataset. The intended language direction is from French to German. **Blue** text denotes the MetricX score in source-based mode, and the **red** text highlights the errors in the translation.

Source (English): In late 2017, Siminoff appeared on shopping television channel QVC.
Reference (Xhosa): Ngasekupheleni kuka-2017, uSiminoff kwitshaneli yokuthenga kwiTV enguQVC.
LLaMAX3-8B-Alpaca (Xhosa): Kwiyeza zokuthenga zesikhathi zesibini ze-2017, uSiminoff wabonakala kwiQVC, isiteshi se-TV seziyobisi. xCOMET: 49.52
LLaMAX3-8B-Alpaca+WALAR (Xhosa): Ngasekupheleni kuka-2017, uSiminoff waboniswa kwinkqubo yentengiso yeQVC, isiteshi seTV. xCOMET: 82.14

Figure 12: Case study of the improvement brought by **WALAR**. The intended language direction is from English to Xhosa. **Blue** text denotes the xCOMET score in reference-based mode.

Source (English): One antibody cocktail, ZMapp, initially showed promise in the field, but formal studies indicated it had less benefit than sought in preventing death.
Reference (Chinese): 一种名为 ZMapp 的“抗体鸡尾酒”（多抗体联合）最初有望在该领域发挥作用，但正式研究表明，它在预防死亡方面的效果有些不尽人意。
LLaMAX3-8B-Alpaca (Chinese): 一种抗体混合物，ZMapp，最初显示出在现场的希望，但正式研究表明它比预期的死亡率降低。 xCOMET: 58.56
LLaMAX3-8B-Alpaca+WALAR (Chinese): 一种抗体混合物，ZMapp，最初显示出希望，但正式研究表明它无法达到预期的防止死亡效果。 xCOMET: 75.04

Figure 13: Case study of the improvement brought by **WALAR**. The intended language direction is English to Chinese. **Blue** text denotes the xCOMET score in reference-based mode.

Source (Chinese): 在 PALM 实验中，ZMapp 用作参照。也就是科学家将其作为基线，把其他三种治疗方法与之作比较。
Reference (Swahili): Katika jaribio la PALM, Zmapp ilitumika kama kidhibiti, kumaanisha wanasayansi waliitumia kama msingi na kulinganisha mat ibabu mengine matatu nayo.
LLaMAX3-8B-Alpaca (Swahili): Katika majaribio ya Palm, ZMapp hutumiwa kama kumbukumbu. Hivyo, wanasayansi wamefanya kulinganisha kati ya matibabu mengine matatu na yake. Kwa kufanya hivyo, wanasayansi wanaweza kuchunguza jinsi matibabu mengine yanavyofanya kazi kwa kulinganisha na ZMapp. xCOMET: 65.48
LLaMAX3-8B-Alpaca+WALAR (Swahili): Katika majaribio ya PALM, ZMapp ilitumiwa kama kipimo cha kulinganisha; yaani, wataalamu walitumia kama kiwango na kulinganisha njia tatu nyingine za matibabu na hiyo. xCOMET: 93.80

Figure 14: Case study of the improvement brought by **WALAR**. The intended language direction is from Chinese to Swahili. **Blue** text denotes the xCOMET score in reference-based mode.

Source (Chinese): 西米诺夫说，2013 年他在《创智赢家》节目中露面后，公司的销售额大增，当时节目组拒绝向这家初创公司投资。
Reference (Swahili): M. Siminoff a déclaré que les ventes ont augmenté après son apparition en 2013 dans un épisode de Shark Tank où le panel de la série a refusé de financer la start-up.
LLaMAX3-8B-Alpaca (French): Westminster a déclaré que depuis son apparition dans l'émission "Entrepreneurs de génie" en 2013, les ventes de sa société ont augmenté considérablement. Cependant, le programme n'a pas investi dans l'entreprise. xCOMET: 15.52
LLaMAX3-8B-Alpaca+WALAR (French): Siminov a déclaré que, après son apparition dans le programme «Entrepreneurs créatifs» en 2013, les ventes de sa société ont considérablement augmenté, mais le programme n'a pas investi dans cette entreprise débutante. xCOMET: 45.24

Figure 15: Case study of the improvement brought by **WALAR**. The intended language direction is from Chinese to French. **Blue** text denotes the xCOMET score in reference-based mode.

Models	en → x	sw → x	tr → x	hi → x	zh → x	ar → x	ru → x
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
LLaMAX3-8B-Alpaca	0.9027	0.8931	0.8975	0.8931	0.8969	0.9018	0.8996
+WALAR	0.9270	0.9224	0.9252	0.9192	0.9258	0.9275	0.9273
Models	x → en	x → sw	x → tr	x → hi	x → zh	x → ar	x → ru
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
LLaMAX3-8B-Alpaca	0.9970	0.8329	0.9719	0.9513	0.9780	0.8329	0.9890
+WALAR	0.9996	0.9619	0.9986	0.9913	0.9887	0.9652	0.9986

Table 9: Complete results for LCR