AfriMTE and AfriCOMET: Enhancing COMET to Embrace Under-resourced African Languages

Jiayi Wang[1], **David Ifeoluwa Adelani**[1,2], **Sweta Agrawal**[3,*], **Marek Masiak**[1,*,†], **Ricardo Rei**[4,5,6], **Eleftheria Briakou**[3], **Marine Carpuat**[3], **Xuanli He**[1], **Sofia Bourhim**[7], **Andiswa Bukula**[8], **Muhidin Mohamed**[9], **Temitayo Olatoye**[10], **Tosin Adewumi**[11], **Hamam Mokayed**[11], **Christine Mwase**[12], **Wangui Kimotho**[2], **Foutse Yuehgoh**[13], **Anuoluwapo Aremu**[2], **Jessica Ojo**[14,2], **Shamsuddeen Hassan Muhammad**[15,2,29], **Salomey Osei**[16,2], **Abdul-Hakeem Omotayo**[17,2], **Chiamaka Chukwuneke**[18,2], **Perez Ogayo**[2], **Oumaima Hourrane**[2], **Salma El Anigri**[19], **Lolwethu Ndolela**[2], **Thabiso Mangwana**[2], **Shafie Abdi Mohamed**[20], **Ayinde Hassan**[21], **Oluwabusayo Olufunke Awoyomi**[22], **Lama Alkhaled**[11], **Sana Al-Azzawi**[11], **Naome A. Etori**[23], **Millicent Ochieng**[24], **Clemencia Siro**[25], **Samuel Njoroge**[26], **Eric Muchiri**[2], **Wangari Kimotho**[27], **Lyse Naomi Wamba Momo**[28], **Daud Abolade**[2], **Simbiat Ajao**[2], **Iyanuoluwa Shode**[2], **Ricky Macharm**[2], **Ruqayya Nasir Iro**[29], **Saheed S. Abdullahi**[30,31], **Stephen E. Moore**[32,33], **Bernard Opoku**[34,2], **Zainab Akinjobi**[35,2], **Abeeb Afolabi**[2], **Nnaemeka Obiefuna**[2], **Onyekachi Raphael Ogbu**[2], **Sam Brian**[2], **Verrah Akinyi Otiende**[36], **Chinedu Emmanuel Mbonu**[37], **Sakayo Toadoum Sari**[38], **Yao Lu**[1], **Pontus Stenetorp**[1]

[1]University College London, UK, [2]Masakhane NLP, [3]University of Maryland, USA,
[4]Unbabel, [5]Instituto Superior Técnico, [6]INESC-ID, [7]ENSIAS, Morocco, [8]SADiLaR, South Africa,
[9]Aston University, UK, [10]University of Eastern Finland, Finland, [11]Luleå University of Technology, Sweden,
[12]Fudan University, China, [13]Conservatoire National des Arts et Métiers, France, [14]Lelapa AI, South Africa,
[15]Imperial College London, UK, [16]University of Deusto, Spain, [17]University of California, USA, [18]Lancaster University, UK,
[19]Mohammed V University, Morocco, [20]Jamhuriya University Of Science and Technology, Somalia,
[21]LAUTECH, Nigeria, [22]The College of Saint Rose, USA, [23]University of Minnesota -Twin Cities, USA,
[24]Microsoft Africa Research Institute, [25]University of Amsterdam, Netherlands, [26]The Technical University of Kenya,
[27]AIMS, Cameroon, [28]KU Leuven, Belgium, [29]HausaNLP, [30]SIAT-CAS, China, [31]Kaduna State University, Nigeria,
[32]University of Cape Coast, Ghana, [33]Ghana NLP, [34]Kwame Nkrumah University of Science and Technology, Ghana,
[35]New Mexico State University, USA, [36]USIU-Africa, [37]UNIZIK, Nigeria, [38]AIMS, Senegal
Corresponding emails: {ucabj45,d.adelani}@ucl.ac.uk,
p.stenetorp@cs.ucl.ac.uk

Abstract

Despite the recent progress on scaling multilingual machine translation (MT) to several under-resourced African languages, accurately measuring this progress remains challenging, since evaluation is often performed on n-gram matching metrics such as BLEU, which typically show a weaker correlation with human judgments. Learned metrics such as COMET have higher correlation; however, the lack of evaluation data with human ratings for under-resourced languages, complexity of annotation guidelines like Multidimensional Quality Metrics (MQM), and limited language coverage of multilingual encoders have hampered their applicability to African languages. In this paper, we address these challenges by creating high-quality human evaluation data with simplified MQM guidelines for error detection and direct assessment (DA) scoring for 13 typologically diverse African languages. Furthermore, we develop AFRICOMET: COMET evaluation metrics for African languages by leveraging DA data from well-resourced languages and an African-centric multilingual encoder (AfroXLM-R) to create the state-of-the-art MT evaluation metrics for African languages with respect to Spearman-rank correlation with human judgments (0.441).

1 Introduction

Recent advances in machine translation (MT) have focused on scaling multilingual translation models and evaluation data to hundreds of languages, including multiple under-resourced languages (Fan et al., 2021a; NLLB-Team et al., 2022; Bapna et al., 2022; Kudugunta et al., 2023). However, measuring the progress made for these under-resourced languages accurately is difficult, since popular n-gram matching metrics, such as BLEU (Papineni

[*] The authors contribute equally to this work and are considered co-third authors.

[†] Currently at the University of Oxford, UK.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5997–6023
June 16-21, 2024

et al., 2002), METEOR (Banerjee and Lavie, 2005), and ChrF (Popović, 2015), fail to capture semantic similarity beyond the lexical level (Zhang et al., 2020; Rei et al., 2020; Sai B et al., 2023). Variants of these metrics have been developed when scaling to various languages such as spBLEU (Fan et al., 2021a), but they often achieve worse correlation to human judgements (Freitag et al., 2022) when compared to embedding-based metrics like BERTScore (Zhang et al., 2020), and learned metrics such as COMET (Rei et al., 2020).

While embedding-based metrics are currently favored for evaluation in MT (Freitag et al., 2022), the application of these metrics to under-resourced languages faces three challenges: (1) lack of high-quality training and evaluation data significantly hampers the development of reliable metrics; (2) the complexity of the Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2014) presents a steep learning curve for non-expert bilingual evaluators, complicating the process of obtaining accurate human assessments; and (3) the limited language coverage of multilingual large language models such as XLM-R (Conneau et al., 2020) restricts their applicability to various low-resource languages (Alabi et al., 2022).

To address these challenges, recent work have utilized the Direct Assessment (DA) scoring annotations (Graham et al., 2013) collected by the organizers of WMT (Rei et al., 2022a) and leveraged the transfer learning capabilities of multilingual encoders to evaluate unseen languages (Rei et al., 2022b; Zerva et al., 2022a). However, the dearth of evaluation data for under-resourced languages such as African languages still remains a significant hurdle in validating these methods. What is worse, as Rei et al. (2020) highlighted, the performance of these approaches is often unpredictable for languages that were not included in the pre-training phase of multilingual language models.

In this paper, we address these challenges by enhancing the state-of-the-art COMET evaluation metric (Rei et al., 2022a) to various under-resourced African languages. To overcome the scarcity of evaluation datasets, we create AfriMTE—a human evaluation dataset focusing on MT adequacy and fluency evaluation for 13 typologically diverse African languages. This is achieved through a participatory research methodology, ensuring a comprehensive and representative data collection process (Nekoto et al., 2020). In addressing the complexities inherent in the MQM framework, we develop a simplified version that aligns with the tenets of Direct Assessment (DA) and is tailored specifically for non-expert evaluators, aiming to augment both usability and accessibility, thereby rendering the evaluation process more accessible to a wider spectrum of evaluators.

Finally, we develop the first COMET model designed for MT evaluation for African languages. Additionally, we introduce the first translation quality estimation (QE) model for African languages, which operates translation quality estimation without requiring reference translations, setting a new benchmark in the QE field (Fan et al., 2019; Specia et al., 2020, 2021; Wang et al., 2021a).

To summarize, our contributions are as follows: (1) we propose simplified MQM evaluation guidelines tailored for non-expert translators; (2) to support the application of our guidelines, we develop a specialized annotation tool; (3) we develop a high-quality human evaluation dataset focusing on machine translation adequacy and fluency for 13 typologically diverse African languages; (4) we establish benchmark systems for MT Evaluation and Quality Estimation by employing transfer learning techniques from existing, well-resourced DA data and utilizing an African-centric multilingual pre-trained language model; (5) to foster ongoing research in the domain of African machine translation evaluation, we will release all evaluation datasets, code, and models publicly.[1]

2 AfriMTE: African Machine Translation Evaluation Dataset

This section details the data and machine translation engines used for annotation, outlines our annotation guidelines and procedure, describes the data quality assurance process, and presents a quantitative analysis of the collected data.

2.1 Dataset and MT Engine

Our annotation work concentrates on the **dev** and **devtest** subsets from the FLORES-200 dataset (NLLB-Team et al., 2022). This is a multi-way parallel dataset designed to enhance MT for low-resource languages. Flores-200 source texts were sampled from English Wikipedia articles and reference translations into target languages were produced by professional translators. We focus on 13 languages pairs (LPs): Darija-French (ary-fra),

[1]The resources will be publicly available at https://github.com/masakhane-io/africomet.

English-Egyptian Arabic (eng-arz), English-French (eng-fra)—a control LP, English-Hausa (eng-hau), English-Igbo (eng-ibo), English-Kikuyu (eng-kik), English-Luo (eng-luo), English-Somali (eng-som), English-Swahili (eng-swh), English-Twi (eng-twi), English-isiXhosa (eng-xho), English-Yoruba (eng-yor), and Yoruba-English (yor-eng). Moreover, we extend our annotation collection to include domain-specific texts from News, TED talks, Movies, and IT domains for English-Yoruba translations, which were established in prior research by Adelani et al. (2021) and Shode et al. (2022), ensuring a comprehensive and domain-varied evaluation. We provide the information of language family groups that our targeted African languages belong to in Table 4 of Appendix A.1.

To acquire MT outputs, we employ two open-source MT engines: NLLB-200 (600M) (NLLB-Team et al., 2022) and M2M-100 (418M) (Fan et al., 2021b). For eng-fra and eng-swh, we generate translations using M2M-100, while for all other LPs, we utilize NLLB-200. This decision is based on the exceptional proficiency of NLLB-200 translations for eng-fra and eng-swh, where our evaluators found them to be almost error-free during the example annotation training phase. While for some LPs such as eng-xho and eng-yor TED talks, despite their overall high translation quality at the sentence level, our evaluators noted minor error at the word level, as shown in Figures 5 and 6 of Appendix A.1. Therefore, we retain the NLLB-200 engine for these languages. The presence of such slight errors provides an opportunity to assess the robustness and sensitivity of our developed metrics in situations with minimal translation errors. When generating translations, we consistently use a beam size of 5 for both engines.

In the FLORES-200 dataset, we sample 270 and 250 sentences respectively from the dev and devtest sets. The sampling reflects the averaged SacreBLEU (Post, 2018) scores for both high-quality and lower-quality translations across 21 language pairs, ensuring a balanced representation of translation effectiveness.[2] Finally, our annotation datasets are structured as triple parallel, comprising <source, machine translation, reference> for all LPs.

2.2 Annotation Guidelines and Tool

This section presents our annotation guidelines and introduces the annotation tool.

[2]Note that our project initially included more LPs, but due to limited evaluators, 13 remained in AfriMTE.

2.2.1 Annotation Guidelines

Recent findings (Freitag et al., 2021a) have indicated that crowd-sourced DA annotations tend to be inconsistent in assessing the quality of high-performing MT systems. This led us to consider adopting the standardized MQM framework (Lommel et al., 2014)—an extensive method for assessing translation quality by defining various error dimensions collected alongside error severity. However, its complex nature presents a learning hurdle for non-expert evaluators, which was recognized during our annotation training phase. Research by Bentivogli et al. (2018) and Chatzikoumi (2020) shows that while DA has traditionally been used for both translation adequacy and fluency, it currently focuses more on adequacy. Moreover, Graham et al. (2013, 2017) suggests employing DA to evaluate both aspects on a 100-point scale. Drawing upon these findings, we propose a simplified MQM guideline focusing on translation adequacy, combining translation accuracy error detection with DA scoring for ease of use by non-expert evaluators. Similarly, we create a distinct MQM guideline for translation fluency, combining translation fluency error detection with DA scoring.

Our evaluators assess translation adequacy and fluency separately, both through a two-dimensional approach: error highlighting and overall DA scoring. In assessing adequacy, evaluators review both the source and translated texts, highlighting errors categorized as "Addition", "Omission", "Mistranslation", and "Untranslated". During the fluency assessment, evaluators focus solely on the translated text, pinpointing errors in "Grammar", "Spelling", "Typography", and "Unintelligible". The specific error definitions are adapted from the original MQM framework.[3]

Upon completing error highlightings, evaluators use the DA guidelines to assign a score between 0 and 100, reflecting the overall quality of adequacy or fluency. We are motivated by the DA+SQM framework (Kocmi et al., 2023) for our DA guidelines, where we additionally bucket scores based on specific levels to reduce subjectivity. Specifically, in these scales, "0" is defined as a "Nonsense/No meaning preserved" translation for adequacy or an "Incomprehensible" translation for fluency, while "100" signifies "Perfect meaning" for adequacy or "Fluent and natural" for fluency. In addition, there are two intermediate score levels within either rat-

[3]https://themqm.org

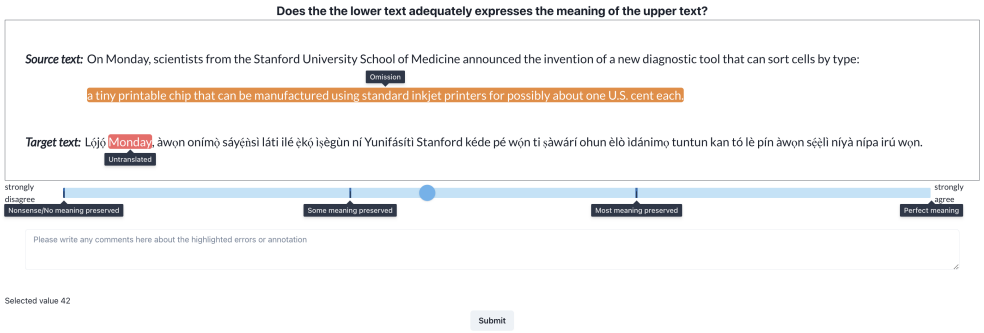

Figure 1: The screenshot of the user interface with an adequacy annotated task comprising the source sentence and its corresponding translation in English-Yoruba.

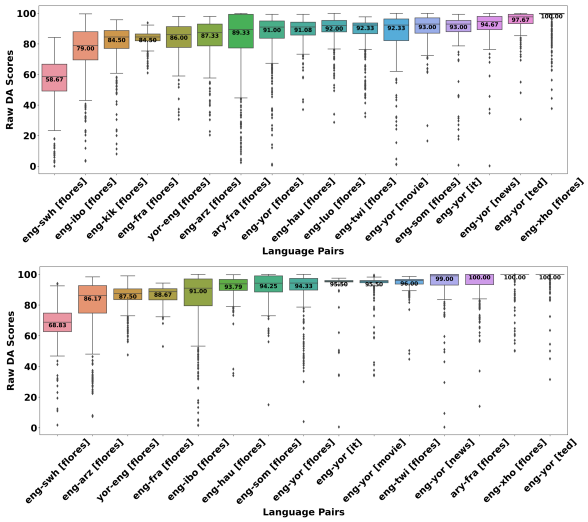

Figure 2: Translation quality of **all** qualified annotated translations as measured by raw DA scores across all language pairs and domains in ascending order, with medians displayed in the plot for **adequacy** (upper) and **fluency** (lower).

ing scale: one at "34" and the another at "67". Details of the adequacy and fluency guidelines are illustrated in Figure 3 and 4 of Appendix A.1, either with two sections: the error highlighting guidelines and the DA scoring guidelines.

2.2.2 Annotation Tool

To collect annotations following our tailored annotation guidelines, we extend an internal annotation tool to suit our needs.[4] Various features have been added, including presenting evaluators with annotation guidelines, adapting the interface to accommodate the error span highlighting and DA scoring functions, and exporting annotations appropriately. The customized tool provides a user-friendly interface designed for machine translation evaluation tasks. A screenshot of the annotation interface is displayed in Figure 1, where every evaluator can work independently.

2.3 Annotation Quality Assurance

We implement a stringent evaluation protocol for each translation, involving a minimum of **two** bilingual native speakers as evaluators, each with a Bachelor's degree or higher. They are encouraged to highlight specific error spans first and then provide a relevant DA score before submission. In preparation, each evaluator annotates 20 examples, and we organize a discussion among the evaluators to review annotations and address any assessment inconsistencies. This preliminary step is designed to familiarize evaluators with the guidelines and the dataset contexts. Some annotators assign low DA scores but lack any corresponding error span highlighting. Hence, in the following data analysis of the correlation between error counts and overall DA scoring, we will exclude such annotations.

Upon completing annotations, we gather data and exclude any with DA score discrepancies exceeding 34 points, as per our guidelines. This threshold is critical for ensuring the reliability of our annotations. To reduce bias among evaluators, we normalize DA scores at the evaluator level to get z-scores, and then average z-scores across evaluators to obtain the final score of each translation. We present the counts of qualified translation annotations within the dev and devtest sets in Table 5 and 6 in Appendix A.1.

To further validate annotation consistency, we apply the inter-annotator agreement (IAA) method (Pavlick and Tetreault, 2016). Each annotation instance is randomly split, with one as Annotator 1 and the average of others as Annotator 2. We compute the Pearson correlation between these two groups, repeating this process 100 times.

[4] https://github.com/marek357/annotation-tool-frontend

The average IAA scores are 0.797 for adequacy and 0.748 for fluency, demonstrating strong consistency among evaluators.

2.4 Quantitative Analysis of Annotations

Overall Translation Quality We show the distributions of raw DA scores across all LPs in Figure 2.[5] Notably, eng-swh translations generated by M2M-100 exhibit the lowest translation adequacy and fluency (median DA: 58.67 and 68.83), whereas eng-xho translations by NLLB-200 score the highest (median DA: 100 for both). Moreover, ary-fra translations have the highest variance in adequacy, while eng-arz and eng-ibo in fluency. The perceived low adequacy and fluency scores observed for the control LP, eng-fra, are consistent with the MT engine employed for it in our study.

Error Counts vs. DA score Equipped with annotations predominantly comprising both overall DA scores and detection of fine-grained error spans, we aim to investigate the correlation between these two aspects. As previously mentioned in Section 2.3, some annotations have low DA scores without error span highlighting. Therefore, we exclude annotations with DA scores under 80 lacking error span highlighting. After this filtration, we present the counts of error words per category and their sentence-level DA scores in Figures 5 and 6 respectively of Appendix A.1 for adequacy and fluency.

Mistranslation is the predominant error impacting adequacy, significantly contributing to lower DA scores. Interestingly, eng-yor Movie translations exhibit a higher incidence of Omission errors, whereas eng-yor IT translations are more prone to Addition errors. Unintelligible is the most common error for fluency except for eng-swh, eng-som, eng-hau. This is consistent in eng-yor domain-specific translations except for the Movie domain.

In order to better understand how word-level errors influence judgments at the sentence level, we calculate and report Pearson, Spearman-rank, and Kendall-rank correlation coefficients between counts within each error category and corresponding scores (raw DA scores and normalized z-scores) in Table 1. These coefficients suggest that Mistranslation and Unintelligible, as the most prevalent error categories for adequacy and fluency, exhibit moderate to high negative correlations with raw DA scores, indicating their significant influence on the sentence-level DA judgements from evaluators. Moreover, for both adequacy and fluency, the total and average error counts per reference translation length negatively correlate with the raw DA scores and the normalized z-scores, further affirming the significance of our simplified MQM guidelines.

[5]We still add domain-specific eng-yor annotations in the plot.

Criteria	Pearson		Spearman		Kendall	
	DA score	Z-score	DA score	Z-score	DA score	Z-score
Mistranslation	-0.478	-0.398	-0.675	-0.544	-0.546	-0.422
Omission	-0.180	-0.196	-0.318	-0.304	-0.263	-0.246
Addition	-0.236	-0.291	-0.207	-0.211	-0.172	-0.172
Untranslated	-0.091	-0.101	-0.156	-0.119	-0.130	-0.097
Total Error	-0.467	-0.479	-0.791	**-0.687**	**-0.640**	**-0.533**
Avg. Error	**-0.490**	**-0.490**	**-0.792**	-0.681	-0.627	-0.516
Grammar	-0.322	-0.191	-0.422	-0.279	-0.355	-0.223
Spelling	-0.042	-0.075	-0.078	-0.103	-0.066	-0.084
Typography	-0.158	-0.257	-0.180	-0.193	-0.153	-0.157
Unintelligible	-0.442	-0.470	-0.466	-0.354	-0.396	-0.286
Total Error	**-0.546**	**-0.577**	**-0.685**	**-0.536**	**-0.576**	**-0.421**
Avg. Error	-0.509	-0.539	**-0.685**	-0.527	-0.568	-0.409

Table 1: Correlations between error counts and sentence-level scores across error categories for **adequacy** (upper) and **fluency** (lower) respectively. "Avg. Error" refers to the average error counts per reference length.

3 AfriCOMET: Benchmark Systems

In this section, we will describe the development of our MT evaluation systems for African languages using AfriMTE. The primary objective of our modeling is to predict the normalized adequacy DA score. Our investigation centers around three key questions: (1) the feasibility of constructing an MT evaluation system that leverages transfer learning from other languages to African languages, (2) the impact of using African language-enhanced pretrained models for MT evaluation systems, and (3) the potential benefits of an additional MT evaluation dataset in African languages for modeling.

Our models are based upon the estimator framework (Rei et al., 2020), as illustrated in Figure 7 of Appendix A.1. In this architecture, the source (src), machine translation (mt), and reference translation (ref) are encoded separately using a multilingual encoder. The resulting word embeddings are passed through a pooling layer to create a sentence embedding for each segment. These sentence embeddings are then combined into a single vector and fed into a feed-forward regressor. The model is trained to minimize the mean squared error. We refer to this as "**single-task learning**" (STL). Furthermore, we adopt a unified approach (Wan et al., 2022), which integrates the tasks of <src, mt>, <mt, ref>, and <src, mt, ref> into one model, feeding all three inputs into the pre-trained model and uniformly distributing weight across the three sentence-level scores for the final score prediction

for MT evaluation. We refer to this as "**multi-task learning**" (MTL).

3.1 Experimental Settings

3.1.1 Datasets

The adequacy **dev** sets in AFRIMTE are employed as validation sets for modeling purposes, while the adequacy **devtest** sets serve as the test sets.

Since 2017, organizers of WMT News translation tasks have been gathering human evaluation using the DA method (Graham et al., 2013). In addition, another large sourced DA annotation set is the MLQE-PE datasets (Fomicheva et al., 2020), typically used in WMT Quality Estimation Shared Tasks (Specia et al., 2020, 2021; Zerva et al., 2022b). We employ these DA datasets as our primary training data,[6] similar to their application in training the COMET metric (COMET22) (Rei et al., 2022a). We label this training data as "**WMT Others**".

Recently, WMT 2022 Large-Scale African Machine Translation Shared Task[7] introduces a DA dataset of 99 source sentences from the FLORES-200 test set (Adelani et al., 2022), covering 46 African language pairs across eight MT engines. Despite its utility, it exhibits two potential limitations: (1) the source context is constrained, consisting of only 99 sentences, and (2) each translation is annotated by a single annotator, raising concerns about the reliability of the assessments. We refer to this dataset as "**WMT African**".

Statistical summaries of the "WMT Others" and "WMT African" datasets are provided in Table 7 and Table 8 respectively in Appendix A.1. Duplicates of <src, mt, ref, DA score> have been excluded. During preprocessing, we also apply z-normalization at the annotator level; to facilitate interpretability and manage the unbounded nature of the quality scores, we apply min-max scaling to the normalized z-scores, adjusting their range to fall between 0 and 1.

3.1.2 Model configurations

In the model setup, we utilize three multilingual pre-trained encoders: XLM-R-L (Conneau et al., 2019), InfoXLM-L (Chi et al., 2020), and an XLM-R-L model adapted to 17 African languages—AfroXLM-R-L (Alabi et al., 2022). Among these, XLM-R-L and InfoXLM-L have been used in the development of COMET22 (Rei et al., 2022a) and CometKiwi (Rei et al., 2022b) metrics for WMT 2022 MT Evaluation and QE Shared Tasks.We provide a detailed overview of language coverage of these three models in Table 9 of Appendix A.1.

We train our models with the open-source COMET codebase.[8] Training of each model is executed on a single NVIDIA A100-SXM4-80GB graphics card, with a configured batch size of 16 and a gradient accumulation across 2 batches. We follow the default settings for other hyper-parameters of the COMET metric.[9]

3.1.3 Evaluation

Pearson, Spearman-rank, and Kendall-rank are widely-used correlation coefficients to assess the correlation between automated and human-annotated scores. Recent findings (Deutsch et al., 2023) indicate that Pearson is complementary to Kendall, and Spearman balances between Pearson's effectiveness in noisy but linear scenarios and Kendall's in ordered but non-linear ones. Thus, we utilize the Spearman-rank correlation coefficient as our primary monitoring metric during model training. For testing, we report all 3 coefficients. To validate the statistical significance of our results, we employ the Perm-Input hypothesis test (Deutsch et al., 2021), conducting 200 re-sampling runs and setting $p = 0.05$. It produces rankings of various automatic metrics. Essentially, for a given test set (in our case, this encompasses translations and their respective assessments), two metrics are juxtaposed using diverse subsets derived from the original test data. The final ranks stem from a significance matrix, which comprises comparisons between all possible pairs of metrics.

3.2 Main Findings

In this section, we will present our experimental results for our investigations around the three key questions mentioned at the beginning of Section 3.

3.2.1 Transfer learning from well-resourced DA data with pre-trained encoders

Initially, we develop our MT evaluation systems that leverage transfer learning from a variety of

[6] We use the DA data from WMT 2017 to 2020 translation tasks and the MLQE-PE data in this work, which can be found in: https://github.com/Unbabel/COMET/tree/master/data. More information is detailed in Freitag et al. (2021b).

[7] https://www.statmt.org/wmt22/large-scale-multilingual-translation-task.html

[8] https://github.com/Unbabel/COMET

[9] Hyper-parameters are configured at https://github.com/Unbabel/COMET/tree/master/configs.

	N-gram Matching		Embedding-based	LLM Prompting	Learned COMET Metrics				
	SacreBLEU	chrf++	BERTScore	GPT-4	**Baseline**	**Single Task** (Ours)			**Multi Task** (Ours)
LP					COMET22	XLM-R-L	InfoXLM-L	AfroXLM-R-L ★	AfroXLM-R-L ★
ary-fra	0.332	0.328	0.351	**0.620**	0.533	0.551	**0.565**	**0.567**	**0.609**
eng-arz	0.324	0.321	0.355	0.509	0.503	0.486	0.488	0.532	**0.600**
eng-fra	0.246	0.280	0.282	**0.536**	**0.489**	**0.510**	0.460	**0.495**	**0.526**
eng-hau	0.200	0.301	0.404	0.378	0.430	0.401	0.334	0.515	**0.620**
eng-ibo	0.339	0.424	0.403	0.271	0.373	0.413	0.377	**0.592**	**0.616**
eng-kik	0.273	**0.295**	0.276	0.269	0.202	0.281	0.249	**0.389**	**0.410**
eng-luo	0.182	0.279	**0.365**	0.246	0.062	0.201	0.241	**0.283**	**0.359**
eng-som	0.161	0.279	0.345	0.281	0.474	0.466	0.420	**0.554**	**0.546**
eng-swh	0.481	0.565	0.701	**0.774**	**0.738**	**0.739**	**0.719**	0.688	**0.733**
eng-twi	**0.204**	**0.178**	0.111	**0.132**	0.096	0.103	**0.112**	**0.157**	0.101
eng-xho	0.090	**0.161**	**0.168**	**0.143**	0.071	0.070	0.059	**0.191**	**0.146**
eng-yor	0.210	0.204	0.250	**0.446**	0.150	0.193	0.191	0.287	0.365
eng-yor (it)	0.295	0.346	**0.421**	**0.447**	0.334	0.256	0.268	0.266	**0.418**
eng-yor (movie)	0.238	0.221	0.303	**0.544**	0.334	0.338	0.364	0.372	0.390
eng-yor (news)	0.114	0.122	0.111	**0.200**	**0.168**	**0.196**	**0.132**	**0.200**	**0.211**
eng-yor (ted)	0.027	0.002	0.091	**0.237**	0.123	0.177	**0.263**	**0.324**	**0.298**
yor-eng	0.308	0.408	0.446	**0.476**	**0.502**	0.460	**0.481**	**0.490**	**0.541**
Avg.	0.237	0.277	0.317	0.383	0.328	0.344	0.337	**0.406**	**0.441**

Table 2: Sentence-level Spearman-rank correlation coefficients for MT evaluation models. For each LP, values in **bold** represent the highest ranking obtained from the Perm-Input hypothesis test (Deutsch et al., 2021). Comprehensive results of this test are detailed in Table 11. Averaged Spearman-rank correlations across LPs are presented in the last row.

well-resourced languages to African languages. We train our models on "WMT Others" and employ the adequacy dev and devtest sets within AFRIMTE as validation and test sets. As outlined in Section 3.1.2, to explore the impact of various multilingual encoders, we conduct experiments based on XLM-R-L, InfoXLM-L, and AfroXLM-R-L for comparison. In our comparison, we benchmark our models against (1) the widely used n-gram matching based evaluation metrics SacreBLEU (Post, 2018) and chrf++ (Popović, 2017), (2) the embedding-based metric, BERTScore (Zhang et al., 2020), (3) LLM Prompting based GPT-4 output with OpenAI API[10] and (4) the learned COMET22 metric (Rei et al., 2022a), which uses the XLM-R-L encoder and also "WMT Others" as training data, but differs in validation, employing additional MQM data for English-German, Chinese-English, and English-Russian from the WMT 2022 News Shared Task.[11]

Results of sentence-level Spearman-rank correlation coefficients are shown in Table 2. Given that "WMT Others" does not include any African language except English, the results of "Learned COMET Metrics" illuminate the effectiveness of various pre-trained multilingual encoders for zero-shot scenarios. Among them, AfroXLM-R-L achieves the highest average result, demonstrating a promising ability to transfer knowledge from well-resourced languages to under-resourced African languages with an African enhanced multilingual encoder. Its performance is further improved with "multi-task learning". We also present Pearson and Kendall-rank correlation coefficient results in Table 10 in Appendix A.1, and the trends observed are consistent with those derived from the Spearman's analysis. Results of Perm-Input hypothesis test for 3 coefficients are illustrated in Table 11, 12 and 13 respectively in Appendix A.1. Both AfroXLM-R-L based systems (STL and MTL) tend to outperform N-gram matching based metrics, BERTScore and COMET22, and show comparable or superior results to GPT-4.

Particularly, our results reveal improvements for eng-ibo and eng-yor (FLORES, News, and TED talks) when we utilize AfroXLM-R-L instead of XLM-R-L as encoder, aligning with their language coverage in Table 9 in Appendix A.1. Additionally, languages initially supported by XLM-R-L, such as eng-hau, eng-som and eng-xho, also experience enhancements with the adoption of AfroXLM-R-L. Interestingly, eng-kik and eng-luo translation evaluations show marked improvements even though Kikuyu and Luo are not explicitly covered by AfroXLM-R-L. Further analysis of correlations across four eng-yor domain-specific datasets show that models trained based on AfroXLM-R-L have the potential to surpass the performance of COMET22, indicating its generalization for different domains despite being trained on the News, Wikipedia and Health domains. For the control LP, eng-fra, our AfroXLM-R-L based systems are

[10] We use the "gpt-4-0613" version, prompting it with the meta-prompt outlined in Figure 8 of Appendix A.1.

[11] https://github.com/google/wmt-mqm-human-evaluation

among the top-performing systems, with distinctions underscored by their bolded rankings. Notably, GPT-4 shows impressive performance in eng-yor and yor-eng MT evaluations.

3.2.2 Impact of an extra African DA dataset

To discuss the potential benefits of an additional MT evaluation dataset in African languages, we conduct experiments based on AfroXLM-R-L across three distinct training data configurations: (1) **"WMT African"**, (2) **"WMT Others"**, and (3) a merged dataset of **"WMT African"** and **"WMT Others"**, which we refer to as **"WMT Combined"**. The STL and MTL results, including Pearson, Spearman-rank, Kendall-rank correlation coefficients, and Perm-Input hypothesis test results, are detailed in Table 14, 15, 16 and 17 respectively in Appendix A.1. Remarkably, "WMT Others" yields higher Spearman-rank and Kendall-rank correlations than "WMT Combined". While "WMT Combined" shows the highest Pearson correlation, it negatively impacts both Spearman-rank and Kendall-rank correlations. Examining all three correlation coefficients and the Perm-Input hypothesis test results reveals that models trained on "WMT Others" and "WMT Combined" significantly outperform the model trained solely on "WMT African". This disparity in performance could be attributed to the limited size and diversity in the context of "WMT African", suggesting its data scarcity issue. In summary, leveraging transfer learning from "WMT Others" based on AfroXLM-R-L proves effective in building African COMET models.

3.3 The benchmark MT evaluation metrics

The AfroXLM-R-L based STL and MTL models, marked with ★ in Table 2, are established as our benchmark MT evaluation systems for African languages. They achieve a Spearman-rank correlation up to 0.441 with human judgments and are named with AfriCOMET-STL and AfriCOMET-MTL.

4 Reference-free QE systems

Utilizing adequacy annotations within AFRIMTE, we are also able to develop reference-free models that predict translation quality in the absence of reference translations, aligning with research advancements in translation quality estimation (QE) (Fan et al., 2019; Ranasinghe et al., 2020; Specia et al., 2020; Wang et al., 2021b,a; Specia et al., 2021; Rei et al., 2022b; Zerva et al., 2022b). Our QE models adhere to the same Estimator architecture as AfriCOMET, but excluding the reference translation from model inputs. Both STL and MTL methods can be applied. However, different from applying MTL in MT evaluation, once the multi-task model is trained, it strictly requires <src, mt> as the input for inference and only generates the corresponding <src, mt> score as its final score.

We choose AfroXLM-R-L and InfoXLM-L for comparison and train our QE models on "WMT Others".[12] These models are validated and evaluated using adequacy dev and devtest sets within AFRIMTE. We benchmark our QE systems against Prompting GPT-4 with meta-prompt as shown in Figure 8 of Appendix A.1 and CometKiwi (Rei et al., 2022b), which is trained on "WMT Others" and leverages InfoXLM-L as its encoder.

When prompting GPT-4, we encounter certain challenges where the API sometimes fails to return the anticipated quality scores, likely attributable to two reasons: the inherent unpredictability of GPT-4's generative capability and its difficulty in identifying some extremely low-resource languages such as Kikuyu, Igbo and Twi in the QE scenario.[13] We find that for eng-kik, eng-ibo, and eng-twi, error rates are markedly higher than the general trend, recorded at 26.2%, 7.5%, and 11.7%, compared to an overall error occurrence below 5% for other LPs. Therefore, for each LP, we implement a missing data imputation approach by assigning the mean of outputs from the remaining valid examples to those that get error responses to ensure consistency and fairness in our evaluation.

QE systems are commonly assessed using Pearson and Spearman-rank correlations as highlighted in (Zerva et al., 2022b). Our results, showcased in Table 3, along with the Perm-Input hypothesis test results in Table 18 in Appendix A.1, reveal the following insights. The InfoXLM-L STL model, trained on "WMT Others", performs on par with CometKiwi under the same encoder configurations. However, the AfroXLM-R-L STL model exhibits significant improvements in both Pearson and Spearman-rank correlations, superior over CometKiwi. Additionally, MTL training further

[12] We follow the hyper-parameter settings at https://github.com/Unbabel/COMET/tree/master/configs, use the same batch size and gradient accumulation, and utilize the same hardware as when training the MT evaluation models.

[13] We re-query the GPT-4 API up to five times for each example, in an effort to obtain successful responses. Despite these efforts, certain instances persist where error responses are encountered even after five attempts.

	LLM Prompting		Learned reference-free QE Metrics							
	GPT-4		**Baseline**		**Single Task** (Ours)				**Multi Task** (Ours)	
			CometKiwi		InfoXLM-L		AfroXLM-R-L ★		AfroXLM-R-L ★	
LP	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
ary-fra	**0.660**	**0.622**	0.517	0.495	0.530	**0.561**	0.475	0.507	**0.610**	**0.534**
eng-arz	0.462	**0.525**	**0.611**	**0.592**	0.562	0.516	0.551	0.516	**0.600**	**0.580**
eng-fra	**0.471**	**0.531**	**0.527**	**0.495**	0.416	**0.484**	**0.418**	**0.478**	**0.483**	**0.531**
eng-hau	0.363	0.284	0.314	0.245	0.382	0.273	**0.652**	0.482	**0.690**	**0.586**
eng-ibo	0.175	0.105	0.205	0.188	0.335	0.334	**0.644**	**0.631**	**0.597**	**0.574**
eng-kik	0.144	0.198	0.277	0.247	0.409	**0.339**	**0.631**	**0.415**	0.437	0.317
eng-luo	0.038	-0.044	0.237	**0.161**	0.142	**0.130**	**0.333**	**0.217**	0.256	**0.174**
eng-som	0.179	0.219	**0.266**	0.357	0.155	0.251	**0.302**	**0.482**	**0.302**	**0.510**
eng-swh	0.693	**0.731**	**0.787**	**0.756**	0.699	0.637	0.644	0.587	0.737	**0.718**
eng-twi	**0.212**	**0.053**	0.097	**0.026**	-0.003	-0.050	**0.290**	**0.061**	**0.279**	**0.060**
eng-xho	**0.254**	**0.119**	0.127	-0.030	0.190	**0.041**	**0.437**	**0.085**	**0.472**	**0.130**
eng-yor	0.339	**0.357**	0.327	0.231	0.489	0.225	**0.738**	**0.392**	0.643	0.280
eng-yor (it)	0.308	0.283	0.375	**0.388**	0.299	0.304	**0.654**	0.318	**0.641**	**0.419**
eng-yor (movie)	0.411	**0.472**	0.151	0.041	0.328	0.240	**0.557**	0.314	0.450	0.311
eng-yor (news)	0.239	**0.126**	0.104	0.078	0.219	0.057	**0.508**	**0.186**	**0.496**	**0.206**
eng-yor (ted)	**0.310**	**0.246**	0.217	**0.289**	0.267	**0.218**	**0.518**	**0.189**	**0.409**	**0.271**
yor-eng	**0.383**	**0.399**	0.070	0.098	-0.007	0.059	0.181	0.208	**0.383**	**0.414**
Avg.	0.332	0.307	0.306	0.274	0.318	0.272	**0.502**	**0.357**	**0.499**	**0.389**

Table 3: Sentence-level correlation coefficients (Pearson, Spearman-rank) for QE models. For each LP, values in **bold** represent the highest ranking obtained from the Perm-Input hypothesis test (Deutsch et al., 2021). The comprehensive results of this test are detailed in Table 18. Averaged correlations across LPs are presented in the last row.

boosts performance in Spearman-rank correlation. These highlight the effectiveness of transfer learning from robust, well-resourced DA data, especially when utilizing AfroXLM-R-L as the pre-trained encoder for the reference-free QE task.

Moreover, when we compare Spearman-rank results in Table 2 and 3, AfroXLM-R-L based QE systems (STL and MTL) outperform GPT-4 by a larger margin than observed in MT evaluation, and the performance gap between QE and MT evaluation systems is larger with GPT-4 $(0.076 = 0.383 - 0.307)$ compared to the AfroXLM-R-L based systems, $(0.049 = 0.406 - 0.357)$ for STL and $(0.052 = 0.441 - 0.389)$ for MTL. This highlights GPT-4's challenges with QE tasks and underscores the superior efficacy of our supervised systems in addressing the inherently cross-lingual nature of QE, diverging from the MT evaluation task. The latter typically involves easier monolingual pattern-matching tasks in comparing machine translations against reference translations.

Finally, we introduce our benchmark QE systems for African MT: the AfroXLM-R-L based STL and MTL models marked with ★ in Table 3, and name them with AfriCOMET-QE-STL and AfriCOMET-QE-MTL.[14]

[14]Please note that AfriCOMET-QE-MTL and AfriCOMET-MTL are identical in training, as both are trained using the same multi-task learning approach.

5 Additional Evaluation

Additional evaluations have been conducted on the generalization of our AfriCOMET and AfriCOMET-QE systems to other datasets. Please refer to Appendix A.2, A.3 and A.4 for details.

6 Conclusion

This study tackles the challenges of enhancing the COMET metric for various under-resourced African languages. We simplify the MQM annotation guidelines for non-expert evaluators, create an MT evaluation dataset, AFRIMTE, covering 13 typologically diverse African languages, and establish benchmark MT evaluation (AFRICOMET) and reference-free QE (AFRICOMET-QE) systems. Our findings show the feasibility of employing transfer learning from well-resourced DA data and an African-centric multilingual pre-trained encoder, AfroXLM-R, for building MT evaluation and QE models for African languages.

Limitations

This work establishes an efficient solution to translation evaluation for under-resourced African languages. It shows that with leveraging a pre-trained model enhanced by under-resourced languages, it is feasible to transfer knowledge from well-resourced to under-resourced languages for the downstream

cross-lingual NLP tasks. However, our current methods are subject to limitations.

Firstly, while using AfroXLM-R-L as a pre-trained encoder enhances the performance of our benchmark systems for certain language pairs such as eng-ibo, eng-kik, eng-luo and eng-yor, this improvement isn't consistent across all LPs. For example, the eng-twi translation evaluation shows no such enhancement and Twi is also not covered by AfroXLM-R-L. Addressing the limited resources and coverage for such under-resourced languages remains a challenge for future work.

Secondly, both our MT evaluation and QE benchmark systems are developed using adequacy annotations within AFRIMTE, mainly drawing inspiration from works by Bentivogli et al. (2018); Chatzikoumi (2020), which suggest that overall DA largely focuses on adequacy. However, upon analyzing the correlations between adequacy and fluency annotations, we have observed a slight negative correlation between total fluency error counts in a translation and its adequacy DA score, with a Pearson correlation coefficient of -0.349. This raises a question: *whether incorporating fluency assessments in developing MT evaluation and QE models could yield any benefit?* Exploring this possibility will be another area for future work.

Thirdly, comparisons of the Spearman-rank results in Table 2 and 3 show significant performance gaps between the AfroXLM-R-L based MT evaluation and reference-free QE systems, even though both employing transfer learning. This disparity likely arises from the tasks' different natures: MT evaluation models are trained with the help of reference inputs, resembling monolingual pattern-recognition tasks that compare machine translations with references. However, the QE task, inherently cross-lingual due to its reference-free nature, highlights the potential need for more training data to bridge this gap. This will be another focus in our future research.

Fourthly, the test datasets in this study, currently limited to translations from a single MT engine per LP, could benefit from diversification. Incorporating outputs from various MT systems into our annotated test data would enrich the spectrum of MT errors observed, significantly enhancing the robustness of metric evaluations. This would be particularly advantageous for developing ranking systems for translation evaluation. However, the expansion is constrained by limited annotation resources, a challenge that is more pronounced for under-resourced African languages within the context of our simplified MQM approach. Despite these challenges, this work's primary goal is to establish reliable metrics for assessing sentence-level translation quality for under-resourced African languages. Our findings demonstrate that our benchmark systems can be used to assess translations from various translation engines.

Ethics Statement

Our work and collection of data has been deeply rooted in the principles of participatory AI research (Nekoto et al., 2020), where the native speakers, most affected by lack of evaluation metrics, are involved throughout the project as stakeholders. They contributed to the data and gave their consent to use this data for the enhancement of COMET models for African languages. Upon the data collected, there is no privacy concern since the source of the data is based on Wikipedia general domain.

Acknowledgments

David Adelani acknowledges the support of DeepMind Academic Fellowship programme. Ricardo Rei is supported by the European Union's Horizon Europe Research and Innovation Actions (UTTER: contract 101070631) and by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI). Pontus Stenetorp would like to acknowledge the helpful proofing feedback from several viewers while finalizing the submission. We are grateful to Prof. Antonios Anastasopoulos from George Mason University for releasing the "WMT African" dataset for our experiments. This work is supported in part by Oracle Cloud credits and related resources provided by Oracle. This work is also supported in part by Microsoft Research via their Accelerate Foundation Models Research Grant. Finally, we are grateful to OpenAI for providing Masakhane with API credits through their Researcher Access API program for the evaluation of the GPT-4 large language models.

References

David Adelani, Md Mahfuz Ibn Alam, Antonios Anastasopoulos, Akshita Bhagia, Marta R. Costa-jussà, Jesse Dodge, Fahim Faisal, Christian Federmann, Natalia Fedorova, Francisco Guzmán, Sergey Koshelev, Jean Maillard, Vukosi Marivate, Jonathan Mbuya,

Alexandre Mourachko, Safiyyah Saleem, Holger Schwenk, and Guillaume Wenzek. 2022. Findings of the WMT'22 shared task on large-scale machine translation evaluation for African languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 773–800, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

David I Adelani, Dana Ruiter, Jesujoba O Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Awokoya, and Cristina España-Bonet. 2021. The effect of domain and diacritics in yor\ub\'a-english neural machine translation. *arXiv preprint arXiv:2103.08647*.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi N. Baljekar, Xavier García, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Z. Chen, Yonghui Wu, and Macduff Hughes. 2022. Building machine translation systems for the next thousand languages. *ArXiv*, abs/2205.03983.

Luisa Bentivogli, Mauro Cettolo, Marcello Federico, and Federmann Christian. 2018. Machine translation human evaluation: an investigation of evaluation based on post-editing and its relation with direct assessment. In *Proceedings of the 15th International Workshop on Spoken Language Translation (IWSLT 2018)*, pages 62–69.

Eirini Chatzikoumi. 2020. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26(2):137–161.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2020. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. *arXiv preprint arXiv:2007.07834*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929.

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

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021b. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1):4839–4886.

Kai Fan, Jiayi Wang, Bo Li, Fengming Zhou, Boxing Chen, and Luo Si. 2019. "bilingual expert" can find translation errors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6367–6374.

Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. 2021. The Eval4NLP shared task on explainable quality estimation: Overview and results. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 165–178, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André FT Martins. 2020. Mlqe-pe: A multilingual quality estimation and post-editing dataset. *arXiv preprint arXiv:2010.04480*.

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

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the wmt21 metrics shared task: Evaluating metrics with expert-based human evaluations on ted and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774.

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.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, et al. 2023. Findings of the 2023 conference on machine translation (wmt23): Llms are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier García, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. *ArXiv*, abs/2309.04662.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. In *Tradumàtica*.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

NLLB-Team, Marta Ruiz Costa-jussà, James Cross, Onur cCelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon L. Spruit, C. Tran, Pierre Yves Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzm'an, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672.

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

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.

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

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copen-

hagen, Denmark. Association for Computational Linguistics.

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

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. Transquest: Translation quality estimation with cross-lingual transformers. *arXiv preprint arXiv:2011.01536*.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022a. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

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. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ananya Sai B, Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra, and Raj Dabre. 2023. IndicMT eval: A dataset to meta-evaluate machine translation metrics for Indian languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14210–14228, Toronto, Canada. Association for Computational Linguistics.

Iyanuoluwa Shode, David Ifeoluwa Adelani, and Anna Feldman. 2022. yosm: A new yoruba sentiment corpus for movie reviews. *arXiv preprint arXiv:2204.09711*.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764, Online. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André FT Martins. 2021. Findings of the wmt 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek F Wong, and Lidia S Chao. 2022. Unite: Unified translation evaluation. *arXiv preprint arXiv:2204.13346*.

Jiayi Wang, Ke Wang, Boxing Chen, Yu Zhao, Weihua Luo, and Yuqi Zhang. 2021a. Qemind: Alibaba's submission to the wmt21 quality estimation shared task. *arXiv preprint arXiv:2112.14890*.

Ke Wang, Yangbin Shi, Jiayi Wang, Yuqi Zhang, Yu Zhao, and Xiaolin Zheng. 2021b. Beyond glass-box features: Uncertainty quantification enhanced quality estimation for neural machine translation. *arXiv preprint arXiv:2109.07141*.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022a. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José GC De Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orašan, Marina Fomicheva, et al. 2022b. Findings of the wmt 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Appendix

A.1 Supplementary materials

The Appendix provides supplementary materials supporting the main paper, including (i) the information of language family groups that our targeted African languages belong to (Table 4) and the statistics of AFRIMTE annotations (Tables 5 and 6), (ii) detailed simplified annotation guidelines (Figures 3 and 4), (iii) distributions of error counts and overall sentence-level DA scores of AFRIMTE annotations (Figures 5 and 6), (iv) the MT evaluation and QE model architecture (Figure 7), (v) meta-prompts for prompting GPT-4 for MT evaluation and QE tasks (Figure 8), (vi) statistical summaries of the "WMT Others" and "WMT African" datasets (Tables 7 and 8), (vii) the overview of language coverage in various pre-trained multilingual models

(Table 9), (viii) the Pearson and Kendall-rank correlation coefficients and the Perm-Input hypothesis test results for MT evaluation models (Tables 10, 11, 12, and 13), (ix) ablation study results for an extra African DA training dataset (Tables 14, 15, 16, and 17), and (x) the Perm-Input hypothesis test results for QE models (Table 18).

A.2 Evaluation on the WMT African DA dataset

Besides evaluating AfriCOMET and AfriCOMET-QE using the adequacy devtest sets within AFRIMTE, we conduct additional assessments on the "WMT African" dataset, despite its potential limitations discussed in Section 3.1.1. These assessments are justified because the "WMT African" dataset is not utilized in the development (training or validation) of the AfriCOMET or AfriCOMET-QE systems. As showcased in Table 19, in the MT evaluation task, AfriCOMET-STL surpasses the cutting-edge COMET22 system across all three correlation coefficients. Meanwhile, AfriCOMET-MTL shows a slight edge over COMET22 in the Pearson correlation coefficient. For QE, both AfriCOMET-QE-STL and AfriCOMET-QE-MTL significantly outperform the state-of-the-art CometKiwi system. These comparisons are fair since all systems are trained using the "WMT Others" dataset. This evaluation further validates the efficacy of our benchmark systems from an additional perspective.

A.3 Evaluation on the WMT 2022 English-Yoruba QE test set

The WMT organizers recently released an English-Yoruba DA dataset, serving as the zero-shot test set in the WMT 2022 Quality Estimation Shared Task. This dataset consists of 1010 DA annotations, prepared using DA guidelines different from ours, as outlined by Fomicheva et al. (2021). The source sentences are sampled from Wikipedia, covering seven topics, and translated into Yoruba using Google Translate, as reported in Zerva et al. (2022b). We evaluate CometKiwi and our benchmark AfriCOMET-QE systems on this dataset. The results, shown in Table 20, demonstrate that our AfriCOMET-QE systems outperform CometKiwi significantly on this English-Yoruba dataset, underscoring the efficacy of our QE approaches even with differently guided DA annotations.

A.4 Generalization Evaluation

Given that our benchmark systems employ the African language-enhanced pre-trained model, AfroXLM-R-L, assessing their generalization capabilities on non-African datasets is crucial. The development of both COMET22 and CometKiwi systems involves using the English-German, English-Russian, and Chinese-English MQM datasets from the WMT 2022 News Domain Translation Shared Task as validation sets, featuring 8959, 8432, and 9750 MQM annotations, respectively. Therefore, testing our benchmark systems on these three datasets is practical to evaluate their generalization in non-African cases. We present the results of correlation coefficients in Table 21. In MT evaluation and QE tasks, AfriCOMET and AfriCOMET-QE exhibit only a slight performance drop compared to COMET22 and CometKiwi systems, respectively, which might be due to the adaptation feasibility of the AfroXLM-R-L pre-trained encoder. This evaluation highlights the sustained generalization capabilities of our benchmark systems.

Language	Family Group
Darija	Afro-Asiatic/Semitic
Egyptian Arabic	Afro-Asiatic/Semitic
English	Indo-European/Germanic/Anglo-Frisian
French	Indo-European/Italic/Romance
Hausa	Afro-Asiatic/Chadic
Igbo	Atlantic-Congo/Volta-Niger
Kikuyu	Atlantic-Congo/Bantu/North-East Bantu
Luo	Nilo-Saharan/Nilotic
Somali	Afro-Asiatic/Cushitic
Swahili	Atlantic-Congo/Bantu/North-East Bantu
Twi	Atlantic-Congo/Kwa
Xhosa	Atlantic-Congo/Bantu/Southern Bantu/Nguni
Yoruba	Atlantic-Congo/Volta-Niger

Table 4: Language family groups that our targeted African languages belong to, according to Wikipedia (https://en.wikipedia.org/wiki/Language_family).

LP	original #	qualified #	dev #	devtest #
ary-fra	520	394	207	187
eng-arz	520	518	268	250
eng-fra	520	515	265	250
eng-hau	520	490	250	240
eng-ibo	520	240	120	120
eng-kik	520	410	208	202
eng-luo	520	499	257	242
eng-som	520	434	208	226
eng-swh	520	352	195	157
eng-twi	520	516	269	247
eng-xho	520	494	251	243
eng-yor	520	484	245	239
eng-yor (it)	250	217	-	217
eng-yor (movie)	270	219	-	219
eng-yor (news)	270	237	-	237
eng-yor (ted)	250	224	-	224
yor-eng	520	439	227	212

Table 5: Counts of qualified **adequacy** annotations for each language pair in dev and devtest sets, with English-Yoruba exclusively as devtest in domain-Specific datasets.

LP	original #	qualified #	dev #	devtest #
ary-fra	520	459	239	220
eng-arz	520	518	268	250
eng-fra	520	459	244	215
eng-hau	520	482	234	248
eng-ibo	520	409	178	231
eng-kik	-	-	-	-
eng-luo	-	-	-	-
eng-som	520	450	224	226
eng-swh	520	376	177	199
eng-twi	520	518	269	249
eng-xho	520	497	250	247
eng-yor	520	495	261	234
eng-yor (it)	250	237	-	237
eng-yor (movie)	270	262	-	262
eng-yor (news)	270	258	-	258
eng-yor (ted)	250	243	-	243
yor-eng	520	500	258	242

Table 6: Counts of qualified **fluency** annotations for each language pair in dev and devtest sets, with English-Yoruba exclusively as devtest in domain-specific datasets.

Adequacy Annotation Guideline

You are asked to compare the meaning of a source segment and its translation. You will be presented with one pair of segments at a time, where a segment may contain one or more sentences. For each pair, you are asked to read the text closely and do the following:

1. Highlight the text spans that convey different meaning in the compared segments. After highlighting a span in the text, you will be asked to select the category that best describes the meaning difference using the following categories:

Source Text:
Omission: *The highlighted span in the source text corresponds to information that does not exist in the translated text.*
Mistranslation: *The highlighted span in the source does not have the exact same meaning as the highlighted span in the translated text.*

Translation Text:
Addition: *The highlighted span in the translation corresponds to information that does not exist in the source text.*
Mistranslation: *The highlighted span in the translation does not have the exact same meaning as the highlighted span in the source segment.*
Untranslated: *The highlighted span in the translation is a copy of the highlighted span in the source segment but should be translated in the target language.*

You can highlight as many spans as needed.

2. Assess the translation adequacy on a continuous scale [0 ~ 100] using the quality levels described below:

[0] Nonsense/No meaning preserved: *Nearly all information is lost between the translation and source.*
[34] Some meaning preserved: *The translation preserves some of the meaning of the source but misses significant parts.*
[67] Most meaning preserved: *The translation retains most of the meaning of the source.*
[100] Perfect meaning: *The meaning of the translation is completely consistent with the source.*

Figure 3: **Adequacy annotation guideline** for error highlighting [the first part] and DA score assignment [the second part].

Fluency Annotation Guideline

You are asked to assess the fluency of a segment. You will be presented with one segment at a time, where a segment may contain one or more sentences. For each segment, you are asked to read it closely and do the following:

1. Highlight the text spans that contain fluency errors. After highlighting a span of text, you will be asked to select the category that best describes the fluency error using the following categories:

Grammar: *The highlighted span corresponds to issues related to the grammar or syntax of the text, other than spelling and orthography.*
Spelling: *The highlighted span corresponds to issues related to spelling of words.*
Typography: *The highlighted span corresponds to issues related to punctuation and diacritics.*
Unintelligible: *The exact nature of the error cannot be determined. Indicates a major break down in fluency.*

You can highlight as many spans as needed.

2. Assess the fluency of the segment on a continuous scale [0 ~ 100] using the quality levels described below:

[0] Incomprehensible: *The translation is completely unintelligible and nonsensical. The text is difficult to understand.*
[34] Poor grammar and disfluent: *The translation contains significant errors in grammar, syntax, and vocabulary that affects the clarity and naturalness of the text.*
[67] Grammatically correct, potentially unnatural: *The translation is grammatically correct but may have some errors in spellings, word choice, or syntax. The language may not be natural.*
[100] Fluent and natural: *The translation contains no grammatical errors, the vocabulary is precise, and the text is easy to read and understand.*

Figure 4: **Fluency annotation guideline** for error highlighting [the first part] and DA score assignment [the second part].

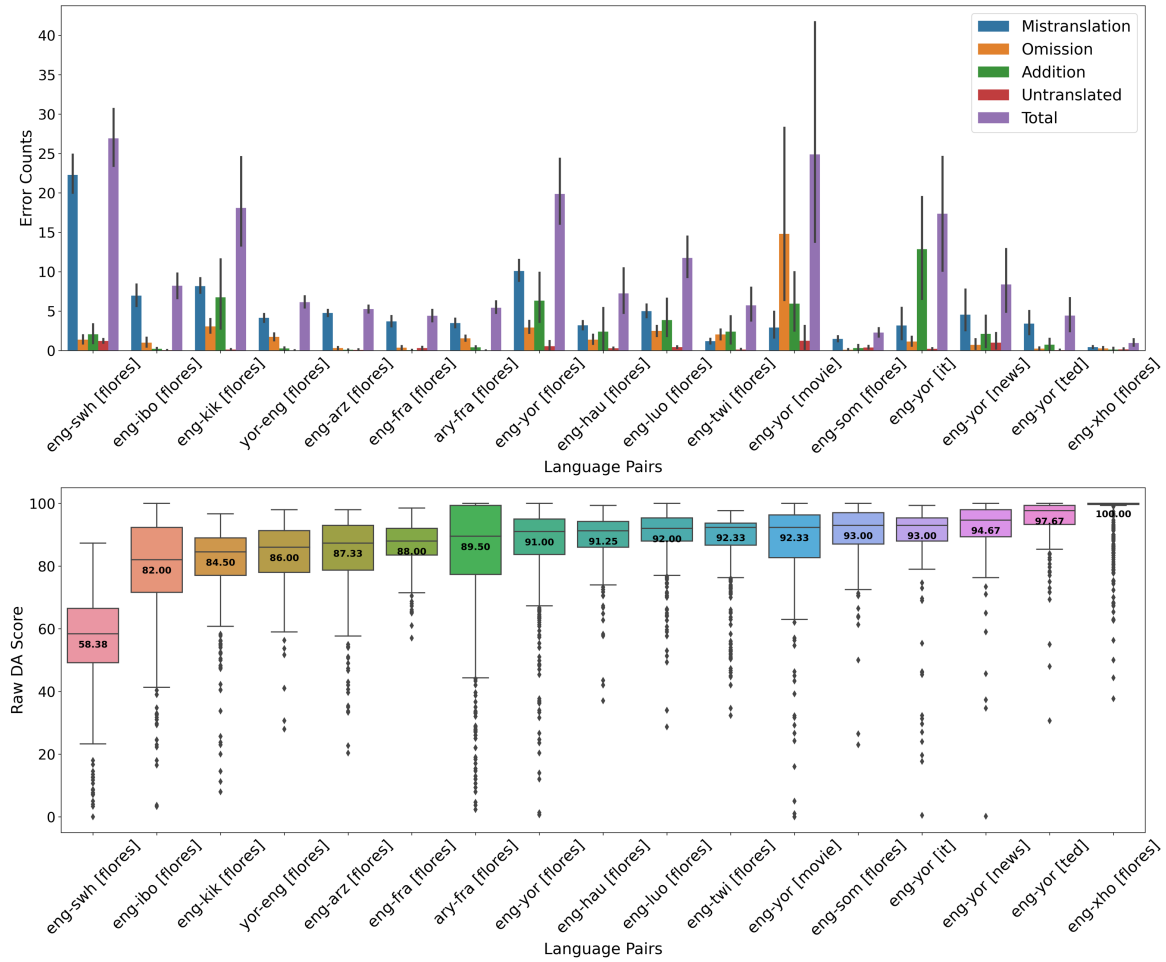

Figure 5: Counts of each error category and sentence-level translation quality measured by DA scores across all language pairs and domains for **adequacy**.

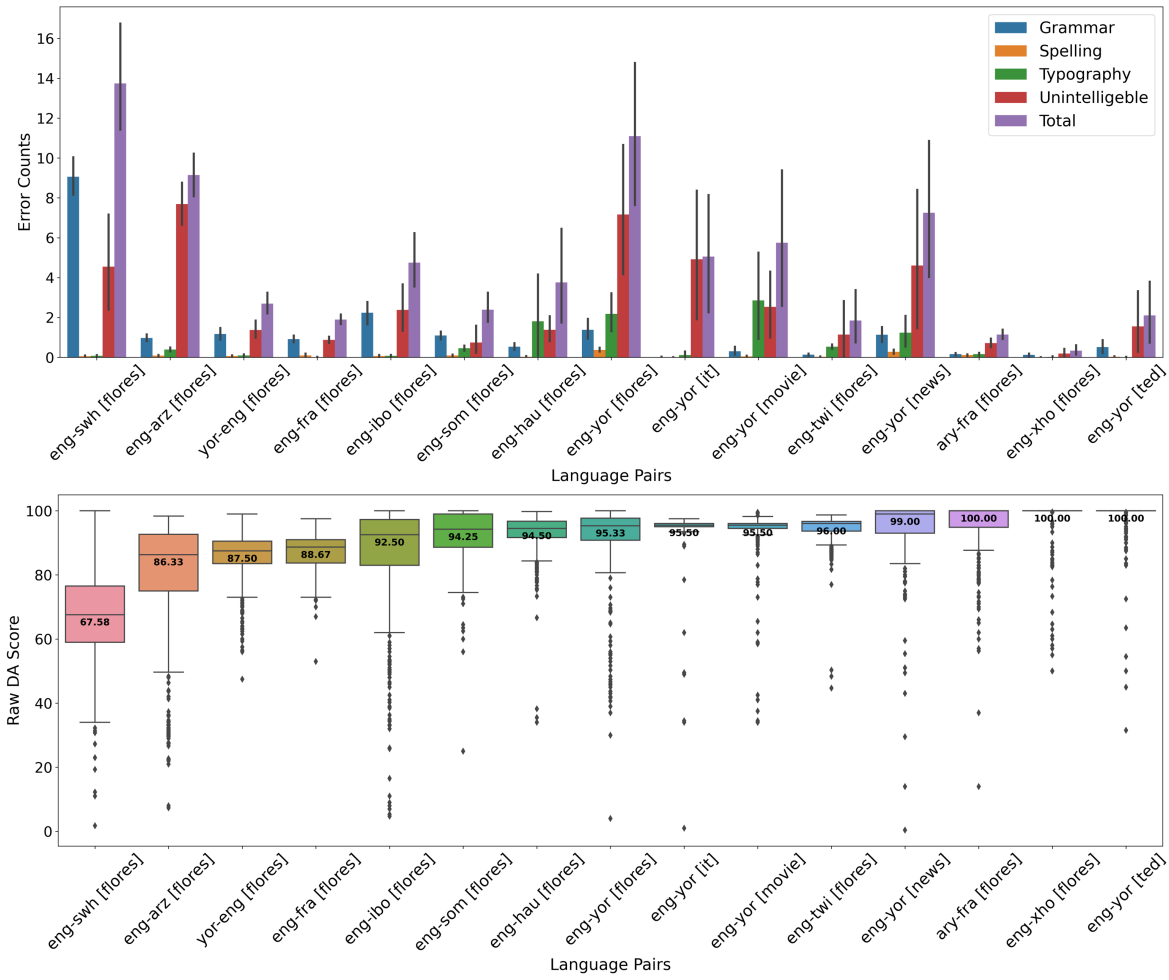

Figure 6: Counts of each error category and sentence-level translation quality measured by DA scores across all language pairs and domains for **fluency**.

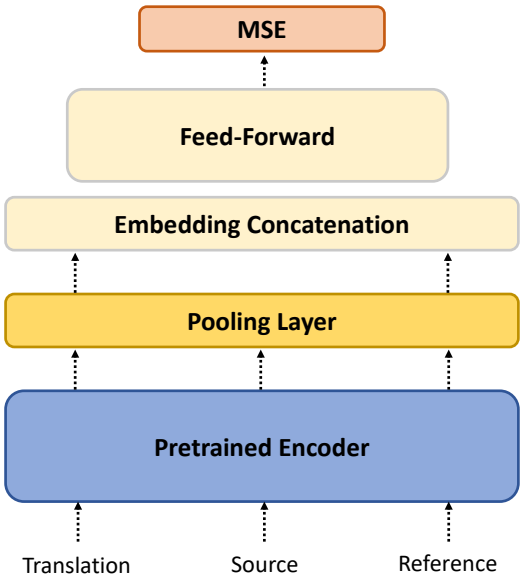

Figure 7: Estimator model architecture. A pre-trained cross-lingual encoder independently encodes the source, translation, and reference. The resulting word embeddings are then passed through a pooling layer to create a sentence embedding for each segment. Then, the corresponding sentence embeddings are combined and concatenated into one single vector, passed to a feed-forward regressor. The entire model is trained by minimizing the Mean Squared Error. Please note that only the source and translation are fed into the pre-trained encoder for training a reference-free QE model.

Meta-Prompt for Prompting GPT4

You are a professional translator. You should assess the machine translation adequacy on a continuous scale [0-100] based on critical points described below:

[0]: Nonsense/No meaning preserved: Nearly all information is lost between the translation and source.
[34]: Some meaning preserved: The translation preserves some of the meaning of the source but misses significant parts.
[67]: Most meaning preserved: The translation retains most of the meaning of the source.
[100]: Perfect meaning: The meaning of the translation is completely consistent with the source.

Note that your score should lie in between two critical points, inclusive of the points themselves.

(for MT evaluation)
Presented below are the source sentence, its machine translation, and the corresponding reference translation:
Source sentence: {source sentence}
Machine translation: {translation sentence}
Reference translation: {reference sentence}

Please assess the above machine translation based on the source sentence and the reference translation. Note that you should only output the final score.

(for Quality Estimation)
Presented below are the source sentence and its machine translation:
Source sentence: {source sentence}
Machine translation: {translation sentence}

Please assess the above machine translation based on the source sentence. Note that you should only output the final score.

Figure 8: Meta prompts utilized in prompting GPT-4 (version: "gpt-4-0613") for MT evaluation and Quality Estimation tasks. Highlights are excluded from the prompts.

LP	Annotation Count	Median	Mean	Std
ces-eng	27847	75.00	69.12	25.18
deu-ces	13804	56.00	53.35	32.97
deu-eng	99183	81.00	73.00	27.06
deu-fra	6691	78.00	71.04	27.44
eng-ces	60937	69.00	62.48	29.09
eng-deu	121420	90.00	80.79	23.2
eng-est	13376	51.00	51.82	29.83
eng-fin	34335	53.00	53.04	30.3
eng-guj	6924	48.50	49.70	28.16
eng-jpn	9578	72.67	68.31	20.45
eng-kaz	8219	57.50	54.16	28.86
eng-lit	8959	60.00	57.40	29.77
eng-lvs	5810	40.00	43.09	29.36
eng-mar	26000	71.75	70.08	10.15
eng-pol	10572	74.00	69.57	22.36
eng-rus	62749	75.00	67.98	27.26
eng-tam	7890	74.00	70.06	19.14
eng-tur	5171	50.00	48.10	33.92
eng-zho	90805	77.00	73.65	20.27
est-eng	29496	70.00	63.48	28.85
fin-eng	46145	75.00	66.29	29.17
fra-deu	3999	83.00	76.13	23.86
guj-eng	9063	58.00	55.70	29.61
jpn-eng	8939	76.00	70.72	24.8
kaz-eng	6789	72.00	64.72	28.09
khm-eng	4722	69.00	61.60	28.01
lit-eng	10315	77.00	70.23	25.31
npi-eng	9000	33.67	37.92	19.51
pol-eng	11816	80.12	76.14	21.62
pbt-eng	4611	70.00	64.14	25.61
ron-eng	9000	76.33	68.76	27.31
rus-eng	79280	84.00	75.38	25.24
sin-eng	9000	50.00	50.45	28.33
tam-eng	7577	72.00	65.45	26.68
tur-eng	30186	71.00	63.51	29.17
zho-eng	126947	79.00	73.37	24.67
Total.	1027155			

Table 7: Statistical summary of **WMT Others** across language pairs: annotation counts, and the median, mean, and standard deviation of the DA scores. Language codes correspond to those specified in FLORES-200 (Goyal et al., 2022).

LP	Annotation Count	Median	Mean	Std
afr-eng	778	78.0	64.14	32.1
afr-ssw	594	68.0	55.32	29.76
amh-eng	594	72.5	60.32	33.4
eng-afr	593	63.0	62.23	30.74
eng-amh	594	55.0	48.37	27.87
eng-hau	592	69.0	58.58	38
eng-ibo	593	71.0	53.59	42.6
eng-kin	594	57.5	53.60	38.32
eng-lug	594	60.0	51.05	38.02
eng-nya	594	81.0	60.44	39.92
eng-orm	594	43.5	43.80	34.17
eng-sna	593	92.0	75.79	36.3
eng-ssw	594	58.0	50.87	33.69
eng-swh	591	85.0	71.13	32.83
eng-tsn	792	80.0	64.48	35.6
eng-xho	594	87.5	61.87	37.56
eng-yor	594	71.0	57.79	35.29
eng-zul	792	84.0	66.19	38.45
fra-lin	594	89.0	70.83	36.68
fra-swh	592	65.0	56.70	30.04
hau-eng	789	83.0	69.94	32.36
hau-ibo	594	48.0	46.74	38.42
ibo-eng	790	82.0	61.38	38.45
ibo-hau	593	69.0	51.78	37.19
ibo-yor	594	52.0	45.48	36.52
kin-eng	590	84.0	65.21	38.05
lin-fra	592	86.5	69.66	36.5
lug-eng	792	42.0	45.95	35.54
nya-eng	594	70.0	58.20	34.64
orm-eng	594	23.0	40.93	39.88
sna-eng	784	91.0	78.65	31.58
som-eng	594	70.0	58.17	34.95
ssw-eng	791	80.0	62.11	40.01
ssw-tsn	594	75.5	66.37	28.07
swh-eng	779	86.0	71.26	33.02
swh-fra	591	83.0	68.68	31.65
swh-lug	594	14.0	30.40	33.41
tsn-eng	791	63.0	54.25	35.24
tsn-tso	594	70.5	63.66	29.68
tso-eng	787	70.0	59.34	36.18
xho-eng	789	85.0	71.72	31.83
xho-zul	594	68.0	49.45	36.56
yor-eng	792	63.0	57.45	33.69
yor-ibo	594	80.0	67.69	33.09
zul-eng	788	90.0	68.47	38.54
zul-sna	593	82.0	64.89	42.39
Total.	30021			

Table 8: Statistical summary of **WMT African** across language pairs: annotation counts, and the median, mean, and standard deviation of DA scores. Language codes correspond to those specified in FLORES-200 (Goyal et al., 2022).

Pre-trained Encoder	Languages Covered	Languages Uncovered
XLM-R-L	English, French, Arabic, Hausa, Somali, Swahili, Xhosa	Igbo, Luo, Kikuyu, Twi, Yoruba
InfoXLM-R-L	English, French, Arabic, Hausa, Somali, Swahili, Xhosa	Igbo, Luo, Kikuyu, Twi, Yoruba
AfroXLM-R-L	English, French, Arabic, Hausa, Igbo, Somali, Swahili, Xhosa, Yoruba	Luo, Kikuyu, Twi

Table 9: Overview of language coverage for XLM-Roberta-Large (XLM-R-L) (Conneau et al., 2019), InfoXLM-Large (InfoXLM-L) (Chi et al., 2020), and AfroXLM-Roberta-Large (AfroXLM-R-L) (Alabi et al., 2022) as utilized in this study.

	N-gram Matching		**Embedding-based**	**LLM Prompting**	**Learned COMET Metric**				
	SacreBLEU	**chrF++**	**BERTScore**	**GPT4**	**Baseline**	**Single Task** (Ours)			**Multi Task** (Ours)
LP					COMET22	XLM-R-L	InfoXLM-L	AfroXLM-R-L ★	AfroXLM-R-L ★
ary-fra	0.307 / 0.234	0.402 / 0.233	0.414 / 0.242	**0.693 / 0.467**	0.584 / 0.379	0.634 / 0.397	0.631 / **0.406**	0.595 / **0.406**	**0.685 / 0.447**
eng-arz	0.241 / 0.222	0.290 / 0.214	0.314 / 0.234	0.454 / **0.379**	0.528 / 0.347	0.533 / 0.339	0.498 / 0.337	0.526 / 0.371	**0.602 / 0.423**
eng-fra	0.268 / 0.171	0.339 / 0.193	0.358 / 0.195	**0.495 / 0.385**	**0.475 / 0.344**	**0.469 / 0.359**	**0.443** / 0.324	**0.515 / 0.351**	**0.522 / 0.372**
eng-hau	0.248 / 0.137	0.445 / 0.206	0.576 / 0.283	**0.664** / 0.278	0.589 / 0.302	0.503 / 0.286	0.473 / 0.229	**0.682** / 0.365	**0.696 / 0.445**
eng-ibo	0.304 / 0.235	0.475 / 0.294	0.365 / 0.292	0.466 / 0.194	0.323 / 0.259	0.386 / 0.288	0.312 / 0.260	**0.551 / 0.435**	**0.649 / 0.445**
eng-kik	0.256 / 0.187	0.406 / **0.202**	**0.498** / 0.188	0.448 / **0.196**	0.434 / 0.139	0.464 / 0.186	0.393 / 0.169	**0.582 / 0.270**	**0.523 / 0.276**
eng-luo	0.182 / 0.122	0.320 / 0.187	**0.429 / 0.250**	0.222 / 0.183	0.203 / 0.039	0.258 / 0.136	0.354 / 0.166	**0.427 / 0.191**	**0.433 / 0.251**
eng-som	0.170 / 0.108	0.317 / 0.196	0.298 / 0.240	**0.485** / 0.205	**0.526** / 0.338	**0.503** / 0.334	**0.465** / 0.297	**0.470 / 0.398**	**0.391 / 0.389**
eng-swh	0.459 / 0.334	0.648 / 0.408	**0.773** / 0.516	**0.768 / 0.604**	**0.779 / 0.560**	**0.771 / 0.567**	**0.775 / 0.546**	0.729 / 0.508	**0.754 / 0.552**
eng-twi	0.185 / **0.137**	0.223 / **0.120**	0.292 / 0.074	**0.456 / 0.096**	**0.378** / 0.064	**0.341** / 0.070	0.274 / **0.078**	**0.396 / 0.104**	0.295 / 0.071
eng-xho	0.124 / 0.072	0.246 / **0.128**	0.306 / **0.132**	**0.433 / 0.117**	0.234 / 0.055	0.202 / 0.054	0.278 / 0.046	**0.473 / 0.150**	**0.465 / 0.115**
eng-yor	0.236 / 0.144	0.355 / 0.143	0.462 / 0.176	**0.674 / 0.334**	0.367 / 0.103	0.329 / 0.131	0.353 / 0.129	0.463 / 0.201	**0.694** / 0.256
eng-yor (it)	0.219 / 0.206	0.411 / 0.244	**0.659 / 0.297**	**0.626 / 0.327**	**0.660** / 0.233	0.558 / 0.177	0.614 / 0.184	0.590 / 0.183	**0.659 / 0.298**
eng-yor (movie)	0.224 / 0.166	0.288 / 0.152	0.430 / 0.213	**0.630 / 0.403**	0.486 / 0.237	0.429 / 0.240	0.503 / 0.256	0.464 / 0.261	0.501 / 0.268
eng-yor (news)	0.207 / 0.081	0.294 / 0.086	0.366 / 0.075	**0.521 / 0.144**	0.395 / **0.118**	0.373 / **0.137**	0.392 / **0.090**	**0.508 / 0.136**	**0.501 / 0.147**
eng-yor (ted)	0.037 / 0.019	0.100 / 0.002	0.284 / 0.062	**0.451 / 0.176**	0.351 / 0.083	0.377 / 0.122	0.449 / **0.185**	**0.539 / 0.224**	0.408 / **0.207**
yor-eng	0.257 / 0.208	0.389 / 0.281	0.425 / 0.308	0.464 / **0.338**	**0.508 / 0.354**	0.452 / 0.323	**0.486** / 0.335	**0.512 / 0.345**	**0.544 / 0.382**
Avg.	0.231 / 0.164	0.350 / 0.193	0.426 / 0.222	0.526 / 0.284	0.460 / 0.233	0.446 / 0.244	0.453 / 0.237	**0.531 / 0.288**	**0.548 / 0.314**

Table 10: Sentence-level Pearson and Kendall-rank correlation coefficients for MT evaluation models. For each LP, values in **bold** represent the highest ranking obtained from the Perm-Input hypothesis test (Deutsch et al., 2021). Comprehensive results of this test are detailed in Table 12 and 13. Averaged Pearson and Kendall-rank correlations across LPs are presented in the last row.

	N-gram Matching		**Embedding-based**	**LLM Prompting**	**Learned COMET Metric**				
	SacreBLEU	chrF++	BERTScore	GPT-4	**Baseline**	**Single Task** (Ours)			**Multi Task** (Ours)
LP					COMET22	XLM-R-L	InfoXLM-L	AfroXLM-R-L ★	AfroXLM-R-L ★
ary-fra	3	3	3	1	2	2	1	1	1
en-arz	4	4	4	2	2	3	3	2	1
en-fra	3	3	3	1	1	1	2	1	1
en-hau	5	4	3	3	2	3	3	2	1
en-ibo	2	2	2	3	2	2	2	1	1
en-kik	2	1	2	2	3	2	2	1	1
en-luo	3	2	1	2	3	2	2	1	1
en-som	5	4	3	3	2	2	2	1	1
en-swh	4	3	2	1	1	1	1	2	1
en-twi	1	1	2	1	2	2	1	1	2
en-xho	2	1	1	1	2	2	2	1	1
en-yor	3	3	2	1	4	3	3	2	2
en-yor (it)	2	2	1	1	2	3	2	2	1
en-yor (movie)	3	3	3	1	2	2	2	2	2
en-yor (news)	2	2	2	1	1	1	1	1	1
en-yor (ted)	3	3	2	1	2	2	1	1	1
yor-eng	3	2	2	1	1	2	1	1	1
Avg.	2.94	2.53	2.24	1.53	2.00	2.06	1.82	**1.35**	**1.18**

Table 11: Detailed rankings from the Perm-Input hypothesis test (Deutsch et al., 2021) of Spearman-rank correlation coefficients corresponding to Table 2. The averaged ranks are presented in the last row.

	N-gram Matching		Embedding-based	LLM Prompting	Learned COMET Metric				
	SacreBLEU	chrf++	BERTScore	GPT-4	**Baseline**	**Single Task** (Ours)			**Multi Task** (Ours)
LP					COMET22	XLM-R-L	InfoXLM-L	AfroXLM-R-L ★	AfroXLM-R-L ★
ary-fra	4	3	3	1	2	2	2	2	1
eng-arz	3	3	3	2	2	2	2	2	1
eng-fra	3	2	2	1	1	1	1	1	1
eng-hau	4	3	2	1	2	3	3	1	1
eng-ibo	3	2	3	2	3	2	3	1	1
eng-kik	3	2	1	2	2	2	3	1	1
eng-luo	4	2	1	3	4	3	2	1	1
eng-som	3	2	2	1	1	1	1	1	1
eng-swh	3	2	1	1	1	1	1	2	1
eng-twi	2	2	2	1	1	1	2	1	2
eng-xho	3	2	2	1	2	3	2	1	1
eng-yor	4	3	2	1	3	4	3	2	1
eng-yor (it)	5	4	1	1	1	3	2	2	1
eng-yor (movie)	4	4	3	1	2	3	2	2	2
eng-yor (news)	3	2	2	1	2	2	2	1	1
eng-yor (ted)	5	5	4	1	3	3	2	1	2
yor-eng	3	2	2	2	1	2	1	1	1
Avg.	3.47	2.65	2.12	**1.35**	1.94	2.24	2.00	**1.35**	**1.18**

Table 12: Detailed rankings from the Perm-Input hypothesis test (Deutsch et al., 2021) of Pearson correlation coefficients corresponding to Table 10. The averaged ranks are presented in the last row.

	N-gram Matching		Embedding-based	LLM Prompting	Learned COMET Metric				
	SacreBLEU	chrf++	BERTScore	GPT-4	**Baseline**	**Single Task** (Ours)			**Multi Task** (Ours)
LP					COMET22	XLM-R-L	InfoXLM-L	AfroXLM-R-L ★	AfroXLM-R-L ★
ary-fra	3	3	3	1	2	2	1	1	1
eng-arz	4	4	4	1	2	2	3	2	1
eng-fra	3	3	3	1	1	1	2	1	1
eng-hau	5	4	3	3	2	3	4	2	1
eng-ibo	2	2	2	3	2	2	2	1	1
eng-kik	2	1	2	1	2	2	2	1	1
eng-luo	3	2	1	2	3	2	2	1	1
eng-som	4	3	3	3	2	2	2	1	1
eng-swh	4	3	2	1	1	1	1	2	1
eng-twi	1	1	2	1	2	2	1	1	2
eng-xho	2	1	1	1	2	2	2	1	1
eng-yor	3	3	3	1	4	3	3	2	2
eng-yor (it)	2	2	1	1	2	3	2	2	1
eng-yor (movie)	3	3	2	1	2	2	2	2	2
eng-yor (news)	2	2	2	1	1	1	1	1	1
eng-yor (ted)	3	3	2	1	2	2	1	1	1
yor-eng	3	2	2	1	1	2	2	1	1
Avg.	2.88	2.47	2.24	1.41	1.94	2.00	1.94	**1.35**	**1.18**

Table 13: Detailed rankings from the Perm-Input hypothesis test (Deutsch et al., 2021) of Kendall-rank correlation coefficients corresponding to Table 10. The averaged ranks are presented in the last row.

	Training Data Settings								
	WMT African			WMT Others			WMT Combined		
LP	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
ary-fra	0.307	0.287	0.201	0.595	0.567	0.406	0.567	0.547	0.388
eng-arz	0.215	0.270	0.177	0.526	0.532	0.371	0.517	0.506	0.351
eng-fra	0.380	0.276	0.190	0.515	0.495	0.351	0.545	0.501	0.355
eng-hau	0.676	0.354	0.240	0.682	0.515	0.365	0.764	0.489	0.342
eng-ibo	0.357	0.406	0.290	0.551	0.592	0.435	0.452	0.562	0.417
eng-kik	0.618	0.256	0.172	0.582	0.389	0.270	0.654	0.368	0.254
eng-luo	0.416	0.255	0.181	0.427	0.283	0.191	0.404	0.275	0.187
eng-som	0.479	0.388	0.271	0.470	0.554	0.398	0.590	0.546	0.390
eng-swh	0.642	0.533	0.373	0.729	0.688	0.508	0.735	0.692	0.515
eng-twi	0.436	0.124	0.082	0.396	0.157	0.104	0.484	0.203	0.139
eng-xho	0.519	0.092	0.072	0.473	0.191	0.150	0.573	0.200	0.155
eng-yor	0.597	0.127	0.083	0.463	0.287	0.201	0.668	0.285	0.202
eng-yor (it)	0.712	0.251	0.172	0.590	0.266	0.183	0.797	0.247	0.172
eng-yor (movie)	0.550	0.274	0.188	0.464	0.372	0.261	0.613	0.349	0.242
eng-yor (news)	0.468	0.066	0.045	0.508	0.200	0.136	0.614	0.204	0.141
eng-yor (ted)	0.404	0.084	0.058	0.539	0.324	0.224	0.608	0.220	0.151
yor-eng	0.406	0.386	0.256	0.512	0.490	0.345	0.511	0.495	0.346
Avg.	0.481	0.261	0.179	0.531	**0.406**	**0.288**	**0.594**	0.393	0.279

Table 14: Correlation coefficients (Pearson, Spearman-rank, Kendall-rank) for MT evaluation models trained with **single-task learning** based on AfroXLM-R-L with varied training data settings. Comprehensive results of the Perm-Input hypothesis test (Deutsch et al., 2021) are detailed in Table 15. The averaged correlation coefficients are presented in the last row.

	Training Data Settings								
	WMT African			WMT Others			WMT Combined		
LP	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
ary-fra	2	2	2	1	1	1	1	1	1
eng-arz	2	3	3	1	1	1	1	2	2
eng-fra	2	2	2	1	1	1	1	1	1
eng-hau	2	2	2	2	1	1	1	1	1
eng-ibo	3	2	2	1	1	1	2	1	1
eng-kik	1	2	2	2	1	1	1	1	1
eng-luo	1	1	1	1	1	1	1	1	1
eng-som	1	2	2	2	1	1	1	1	1
eng-swh	2	2	2	1	1	1	1	1	1
eng-twi	1	2	2	2	1	2	1	1	1
eng-xho	1	2	2	2	1	1	1	1	1
eng-yor	2	2	2	3	1	1	1	1	1
eng-yor (it)	2	1	1	3	1	1	1	1	1
eng-yor (movie)	2	2	2	3	1	1	1	1	1
eng-yor (news)	2	2	2	1	1	1	1	1	1
eng-yor (ted)	2	3	3	1	1	1	1	2	2
yor-eng	2	2	2	1	1	1	1	1	1
Avg.	1.76	2.00	2.00	1.65	**1.00**	**1.06**	**1.06**	1.12	1.12

Table 15: Detailed rankings from the Perm-Input hypothesis test (Deutsch et al., 2021) of Pearson, Spearman-rank and Kendall-rank correlation coefficients for MT evaluation models trained with **single-task learning** based on AfroXLM-Roberta-Large with varied training data settings, corresponding to Table 14. The averaged ranks are presented in the last row.

	Training Data Settings								
	WMT African			WMT Others			WMT Combined		
LP	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
ary-fra	0.262	0.242	0.174	0.685	0.609	0.447	0.677	0.599	0.433
eng-arz	0.293	0.276	0.186	0.602	0.600	0.423	0.600	0.586	0.412
eng-fra	0.142	0.032	0.019	0.522	0.526	0.372	0.486	0.500	0.351
eng-hau	0.530	0.090	0.064	0.696	0.620	0.445	0.774	0.579	0.410
eng-ibo	0.124	0.196	0.140	0.649	0.616	0.445	0.621	0.507	0.364
eng-kik	0.519	0.233	0.161	0.523	0.410	0.276	0.630	0.332	0.225
eng-luo	0.320	0.270	0.181	0.433	0.359	0.251	0.460	0.370	0.252
eng-som	0.280	0.306	0.208	0.391	0.546	0.389	0.426	0.576	0.408
eng-swh	0.543	0.380	0.258	0.754	0.733	0.552	0.752	0.716	0.534
eng-twi	0.438	0.170	0.115	0.295	0.101	0.071	0.467	0.133	0.092
eng-xho	0.505	0.022	0.016	0.465	0.146	0.115	0.663	0.144	0.113
eng-yor (flores)	0.716	0.186	0.126	0.694	0.365	0.256	0.811	0.323	0.227
eng-yor (it)	0.741	0.298	0.208	0.659	0.418	0.298	0.817	0.261	0.255
eng-yor (movie)	0.482	0.092	0.060	0.501	0.390	0.268	0.572	0.314	0.214
eng-yor (news)	0.435	0.018	0.012	0.501	0.211	0.147	0.615	0.115	0.077
eng-yor (ted)	0.384	0.035	0.027	0.408	0.298	0.207	0.553	0.179	0.123
yor-eng	0.292	0.287	0.193	0.544	0.541	0.382	0.535	0.552	0.389
Avg.	0.412	0.184	0.126	0.548	**0.441**	**0.314**	**0.615**	0.399	0.287

Table 16: Correlation coefficients (Pearson, Spearman-rank, Kendall-rank) for MT evaluation models trained with **multi-task learning** based on AfroXLM-R-L with varied training data settings. Comprehensive results of the Perm-Input hypothesis test (Deutsch et al., 2021) are detailed in Table 17. The averaged correlation coefficients are presented in the last row.

	Training Data Settings								
	WMT African			WMT Others			WMT Combined		
LP	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
ary-fra	2	2	2	1	1	1	1	1	1
eng-arz	2	2	2	1	1	1	1	1	1
eng-fra	3	3	2	1	1	1	2	2	1
eng-hau	2	2	3	1	1	1	1	1	2
eng-ibo	2	3	3	1	1	1	1	2	2
eng-kik	2	3	2	2	1	1	1	2	2
eng-luo	2	2	2	1	1	1	1	1	1
eng-som	3	3	2	2	2	1	1	1	1
eng-swh	2	2	2	1	1	1	1	1	1
eng-twi	1	1	1	2	1	1	1	1	1
eng-xho	2	2	2	2	1	1	1	1	1
eng-yor	2	2	2	2	1	1	1	1	1
eng-yor (it)	2	2	2	3	1	1	1	1	2
eng-yor (movie)	2	3	3	2	1	1	1	2	2
eng-yor (news)	2	2	2	2	1	1	1	2	2
eng-yor (ted)	2	3	3	1	1	1	1	2	2
yor-eng	2	2	2	1	1	1	1	1	1
Avg.	2.06	2.29	2.18	1.53	**1.06**	**1.00**	**1.06**	1.35	1.41

Table 17: Detailed rankings from the Perm-Input hypothesis test (Deutsch et al., 2021) of Pearson, Spearman-rank and Kendall-rank correlation coefficients for MT evaluation models trained with **multi-task learning** based on AfroXLM-Roberta-Large with varied training data settings, corresponding to Table 16. The averaged ranks are presented in the last row.

	LLM Prompting		Learned refence-free QE Metric							
	GPT4		**Baseline**		**Single Task** (Ours)				**Multi Task** (Ours)	
			CometKiwi		InfoXLM-L		AfroXLM-R-L ★		AfroXLM-R-L ★	
LP	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
ary-fra	1	1	2	2	2	1	3	2	1	1
eng-arz	2	1	1	1	2	2	2	2	1	1
eng-fra	1	1	1	1	2	1	1	1	1	1
eng-hau	2	3	2	3	2	3	1	2	1	1
eng-ibo	3	3	3	3	2	2	1	1	1	1
eng-kik	4	2	3	2	2	1	1	1	2	2
eng-luo	3	2	2	1	3	1	1	1	2	1
eng-som	2	3	1	2	2	3	1	1	1	1
eng-swh	2	1	1	1	2	2	3	2	2	1
eng-twi	1	1	2	1	2	2	1	1	1	1
eng-xho	1	1	2	2	2	1	1	1	1	1
eng-yor	4	1	4	2	3	2	1	1	2	2
eng-yor (it)	2	2	2	1	2	2	1	2	1	1
eng-yor (movie)	2	1	4	3	3	2	1	2	2	2
eng-yor (news)	2	1	2	2	2	2	1	1	1	1
eng-yor (ted)	1	1	2	1	2	1	1	1	1	1
yor-eng	1	1	3	2	3	3	2	2	1	1
Avg.	2.00	1.53	2.18	1.76	2.24	1.82	**1.35**	**1.41**	**1.29**	**1.18**

Table 18: Detailed rankings from the Perm-Input hypothesis test (Deutsch et al., 2021) of Pearson and Spearman-rank correlation coefficients corresponding to Table 3. The averaged ranks are presented in the last row.

MT Evaluation			
MT Evaluation System	Pearson	Spearman	Kendall
COMET22 (Rei et al., 2022a)	0.578	0.482	0.332
AfriCOMET-STL (Ours)	**0.618**	**0.507**	**0.351**
AfriCOMET-MTL (Ours)	0.591	0.486	0.333
Quality Estimation			
QE System	Pearson	Spearman	Kendall
CometKiwi (Rei et al., 2022b)	0.242	0.219	-
AfriCOMET-QE-STL (Ours)	0.552	0.413	-
AfriCOMET-QE-MTL (Ours)	**0.558**	**0.445**	-

Table 19: Performance of COMET22, AfriCOMET, CometKiwi and AfriCOMET-QE on the "**WMT African**" dataset, a human evaluation set from the WMT 2022 shared task: "Large-Scale Machine Translation Evaluation for African Languages" (Adelani et al., 2022). Results are reported in terms of correlation coefficients: Pearson, Spearman-rank, and Kendall-rank for MT evaluation; Pearson and Spearman-rank for QE. MT evaluation systems are evaluated using the source, the machine translation, and the reference as model inputs, while QE systems are assessed relying only on the source and the machine translation. Correlations are calculated between human-annotated DA scores and automatic scores.

Quality Estimation		
QE System	Pearson	Spearman
CometKiwi (Rei et al., 2022b)	0.153	0.118
AfriCOMET-QE-STL (Ours)	0.461	0.482
AfriCOMET-QE-MTL (Ours)	**0.485**	**0.495**

Table 20: Performance of CometKiwi and our benchmark AfriCOMET-QE systems on the English-Yoruba test set (https://github.com/WMT-QE-Task/wmt-qe-2022-data/tree/main/test_data-gold_labels/task1_da/en-yo) from the WMT 2022 Quality Estimation Shared Task (Zerva et al., 2022b). This dataset includes 1010 DA annotations. Results are reported in terms of Pearson and Spearman-rank correlation coefficients. All metrics are trained on the "WMT Others" dataset, and they are evaluated with source and machine translation as model inputs. Correlations are calculated between human-annotated DA scores and automatic scores.

	MT Evaluation									Quality Estimation					
	COMET22			**AfriCOMET-STL** (Ours)			**AfriCOMET-MTL** (Ours)			**CometKiwi**		**AfriCOMET-QE-STL** (Ours)		**AfriCOMET-QE-MTL** (Ours)	
LP	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
eng-deu	0.312	0.319	0.244	0.263	0.277	0.211	0.265	0.286	0.219	0.254	0.273	0.264	0.256	0.228	0.247
eng-rus	0.361	0.370	0.286	0.344	0.341	0.264	0.381	0.380	0.295	0.357	0.360	0.326	0.337	0.336	0.358
zho-eng	0.428	0.490	0.357	0.427	0.487	0.355	0.420	0.479	0.348	0.362	0.423	0.370	0.421	0.367	0.431
Avg.	**0.367**	**0.393**	**0.296**	0.345	0.368	0.277	0.355	0.382	0.287	**0.324**	**0.352**	0.320	0.338	0.310	0.345

Table 21: Generalization assessments: performance of COMET22 and AfriCOMET for MT evaluation tasks, and performance of CometKiwi and AfriCOMET-QE for QE tasks, on the English-German (eng-deu), English-Russian (eng-rus), and Chinese-English (zho-eng) MQM datasets from the WMT 2022 News Domain Translation Shared Task (https://github.com/google/wmt-mqm-human-evaluation). These three datasets serve as validation sets for COMET22 and CometKiwi, while remaining unseen in either training or validation for AfriCOMET and AfriCOMET-QE.