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

Jiayi Wang[1], David Ifeoluwa Adelani[1*], Sweta Agrawal[2], Ricardo Rei[3,4,5], Eleftheria Briakou[2] Marine Carpuat[2], Marek Masiak[6], Xuanli He[1], Sofia Bourhim[7], Andiswa Bukula[8] Muhidin Mohamed[9], Temitayo Olatoye[10], Hamam Mokayede[11], Christine Mwase[12], Wangui Kimotho*, Foutse Yuehgoh[13], Anuoluwapo Aremu*, Jessica Ojo[14*], Shamsuddeen Hassan Muhammad[15*], Salomey Osei[16*], Abdul-Hakeem Omotayo[17*], Chiamaka Chukwuneke[18*], Perez Ogayo*, Oumaima Hourrane*, Salma El Anigri[19] Lolwethu Ndolela*, Thabiso Mangwana*, 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*, Wangari Kimotho[27], Lyse Naomi Wamba Momo[28], Daud Abolade*, Simbiat Ajao*, Tosin Adewumi[11], Iyanuoluwa Shode*, Ricky Macharm*, Ruqayya Nasir Iro[29], Saheed S. Abdullahi[30,31], Stephen E. Moore[32,33], Bernard Opoku[34*], Zainab Akinjobi[35*], Abeeb Afolabi*, Nnaemeka Obiefuna*, Onyekachi Raphael Ogbu*, Sam Brian*, Verrah Akinyi[36] Chinedu Emmanuel Mbonu[37], Pontus Stenetorp[1]

* Masakhane NLP, [1]University College London, United Kingdom, [2]University of Maryland, USA, [3]Unbabel
[4] Instituto Superior Técnico, [5]INESC-ID, [6]University of Oxford, United Kingdom, [7]ENSIAS, Morocco,
[8]SADiLaR, South Africa, [9] Aston University, United Kingdom, [10]University of Eastern Finland, Finland,
[11]Luleå University of Technology, Sweden, [12] Fudan University, China [13]Devinci Research Center, France,
[14]Lelapa AI, South Africa [15] Bayero University, Nigeria, [16]University of Deusto, Spain
[17] University of California, USA, [18]Lancaster University, United Kingdom [19]Mohammed V University, Morocco,
[20] Jamhuriya University Of Science and Technology, Somalia, [21]IThe College of Saint Rose, USA [22]LAUTECH, Nigeria
[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]UCAS, 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.

Abstract

Despite the progress we have recorded in scaling multilingual machine translation (MT) models and evaluation data to several under-resourced African languages, it is difficult to measure accurately the progress we have made on these languages because evaluation is often performed on n-gram matching metrics like BLEU that often have worse correlation with human judgements. Embedding-based metrics like COMET correlates better, however, 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 a high-quality human evaluation data with a simplified MQM guideline for error-span annotation and direct assessment (DA) scoring for 13 typologically diverse African languages. Furthermore, we develop AFRICOMET—a COMET evaluation metric for African languages by leveraging DA training data from high-resource languages and African-centric multilingual encoder (AfroXLM-Roberta) to create the state-of-the-art evaluation metric for African languages MT with respect to Spearman-rank correlation with human judgements ($+0.406$).

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, it is difficult to measure accurately the progress we have made on these under-resourced languages because popular n-gram matching metrics like BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ChrF (Popović, 2015) fail to capture se-

mantic similarity beyond the lexical level. Variant of these metrics have been developed when scaling to many languages like 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), BLEURT (Sellam et al., 2020) and COMET (Rei et al., 2020).

While embedding-based metrics are currently favored for evaluation in machine translation (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 for these languages 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-Roberta (Conneau et al., 2020), restricts their applicability to various low-resource languages, as discussed in Alabi et al. (2022).

Addressing these challenges, recent works have utilized the Direct Assessment (DA) scoring annotations (Graham et al., 2013) collected by the organisers 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's 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 the challenges of empowering the COMET evaluation metric—the current best 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 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). To address the complexities associated with the MQM guidelines for MT adequacy, we introduced a more simplified version harmonized with the principles of Direct Assessment, specifically adapted to suit non-expert evaluators. This adaptation aims to enhance usability and accessibility, making the evaluation process more approachable for a broader range of users. Finally, we tackle the challenge of empowering the COMET evaluation metric, recognized as the foremost metric in the field according to Rei et al. (2022a), for a range of under-resourced African languages. We develop the first COMET model specifically designed for African languages, which were previously uncovered by the available state-of-the-art COMET models.

To summarise, our contributions are as follows:

1. **Development of a simplified MQM Guideline**: We propose a simplified Multidimensional Quality Metrics (MQM) framework tailored for non-expert translators. This initiative aims to standardize and elevate the quality of human evaluation of machine translation (MT) models.

2. **Creation of African-centric Human Evaluation Datasets**: We develop high-quality human evaluation datasets focusing on machine translation adequacy for 13 typologically-diverse African languages. This endeavor enriches the resources available for evaluating MT models in underrepresented languages.

3. **Benchmark Models for COMET**: We establish benchmark COMET systems for African languages by employing transfer learning techniques from existing, well-resourced Direct Assessment data and utilizing multilingual pre-trained language models.

4. **Open Access to Resources**: In our commitment to fostering ongoing research in the domain of African machine translation evaluation, we are releasing all evaluation datasets, code, and models publicly [1]. This move ensures that researchers and practitioners in the field can easily access and leverage these resources for future advancements.

2 AFRIMTE: African Machine Translation Evaluation Dataset

The section outlines the data and machine translation engines utilized for annotation, and details the design of our annotation guidelines, the conduct

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

of the annotation procedure, the methods implemented to ensure the quality of the data collected, the quantitative analysis conducted on the annotations we collected.

2.1 Dataset and MT Engine

Our annotation efforts focus on the **dev** and **devtest** datasets from the FLORES-200 dataset (NLLB-Team et al., 2022), covering 13 language pairs: Darija-French (ary-fra), English-Egyptian Arabic (eng-arz), English-French (eng-fra), 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-Xhosa (eng-xho), English-Yoruba (eng-yor), and Yoruba-English (yor-eng) [2]. Moreover, to assess the performance of machine translation evaluation across diverse domains, we extend our annotation collection to include texts from News, Ted talk, Movie, and IT domains for English-Yoruba translation. This aspect of our study follows the methodologies established in prior research by Adelani et al. (2021) and Shode et al. (2022), ensuring a comprehensive and domain-varied evaluation.

To acquire machine translation outputs, we employed two open-source translation engines: NLLB-200 (NLLB-Team et al., 2022) and M2M-100 (Fan et al., 2021b). For the language pairs English-French and English-Swahili, we generated translations using M2M-100, while for all other language pairs, we utilized NLLB-200. This decision was informed by the notably high quality of the NLLB-200 translations for English-French and English-Swahili, which were so proficient that our evaluators found minimal errors. However, for certain languages, such as English-Xhosa, we continued to use the high-quality translations provided by NLLB-200. This scenario of near-flawless machine translation outputs from NLLB-200 offers a valuable context for testing the robustness and sensitivity of our evaluation methods in scenarios where translation errors are minimal.

2.2 Annotation Guidelines and Tool

Recent findings (Freitag et al., 2021) have indicated that DA annotations sourced from non-professional crowd annotators tend to be inconsistent and unreliable for assessing the quality of high-performing machine translation (MT) systems. This let us to consider adopting the standardized MQM guideline framework. This framework provides an extensive methodology for assessing translation errors, by defining over various error dimensions, collected alongside the severity and priority of translation errors. However, the complex nature of the MQM framework presents a significant learning hurdle for non-expert evaluators, which was recognized during the training phase of annotation. To address this issue, we develop a novel approach that combines the strengths of MQM and DA annotations. We propose a simplified version of the MQM guideline, designed to be more accessible for non-expert evaluators. This approach involves evaluators identifying error spans using the simplified MQM framework before assigning DA scores. This integration of MQM-based error detection aims to enhance the quality and accuracy of DA annotations, while making the process more manageable for non-expert evaluators.

In our study, we collect DA score annotations by closely adhering to the human evaluation guidelines outlined in Adelani et al. (2022). Additionally, leveraging our simplified Multidimensional MQM guidelines for machine translation adequacy, we prompt evaluators to conduct a detailed error span highlighting. This step precedes the assignment of DA scores, allowing for a more nuanced and precise evaluation process.

2.2.1 Annotation Guidelines

We guide our evaluators to evaluate the adequacy of translations along two dimensions—error highlighting and overall Direct Assessment (DA) score assignment. The process involves presenting the evaluators with both the source text and the machine-translated output. They are instructed to identify and highlight text spans that contained errors in the source text, such as omissions and mistranslations, as well as in the target text, including additions, mistranslations, and untranslated segments. These specific error categories are derived and adapted from the original MQM framework [3], providing a structured approach to error identification and assessment. Subsequently, evaluators were instructed to assign a value between 0 and 100 to indicate the extent to which the original meaning was preserved in the translation. In this scale, "0" is defined as "Nonsense/No meaning preserved," while "100" signifies "Perfect meaning." Echoing the insights

[2] We follow the language codes used in FLORES-200, https://github.com/facebookresearch/flores/tree/main/flores200

[3] https://themqm.org/

Annotation Guidelines

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 1: The annotation guidelines for error span highlighting [the first part] and DA score assignment [the second part].

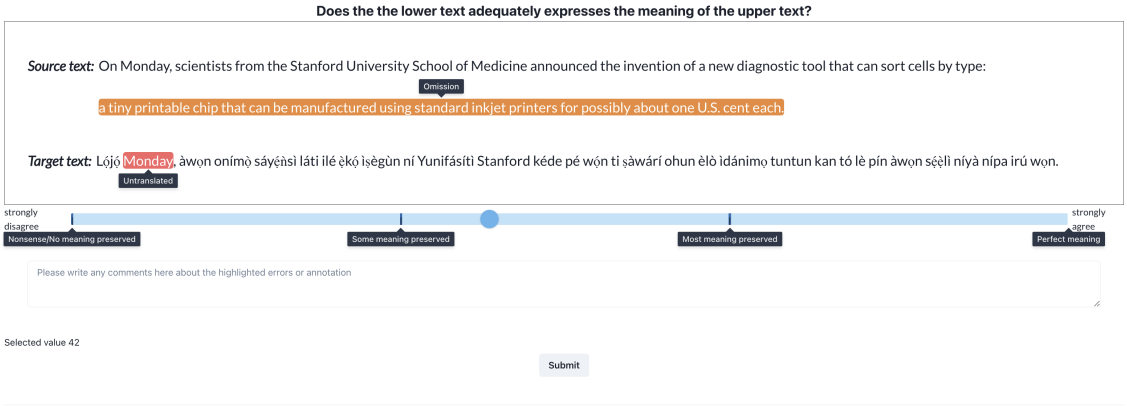

Figure 2: The screenshot of the user interface with an annotated task comprising the source sentence and its corresponding translation in English-Yoruba.

from Adelani et al. (2022), we acknowledged that such an evaluation scale is inherently subjective. To mitigate this subjectivity and provide clearer benchmarks, we established two intermediate levels within the rating scale: one at 33, labeled as "Some meaning preserved," and another at 67, designated as "Most meaning preserved". The guidelines are shown in Figure 1. The first part is our simplified MQM guidelines for error span highlighting, and the second part is the guideline for DA score assignment.

2.2.2 Annotation Tool

For the purpose of collecting annotations in accordance with our tailored annotation guidelines, we extended Annopedia, [4] an open-source tool [5] to suit the needs of MT evaluation for adequacy. The customized tool provides a user-friendly interface specifically designed for machine translation evaluation tasks. It is adept at accommodating both the Direct Assessment (DA) score annotation guidelines and our simplified Multidimensional Quality Metrics (MQM) annotation guidelines. Evaluators can intuitively highlight fine-grained error spans in presented texts and assign DA scores on an assessment scale bar, and they will be informed of the chosen DA score once they finish annotations for a specific translation. A screenshot of the interface is

[4] https://github.com/marek357/best-dissertation-frontend

[5] Tool is accessible online at https://mt.annopedia.marekmasiak.tech

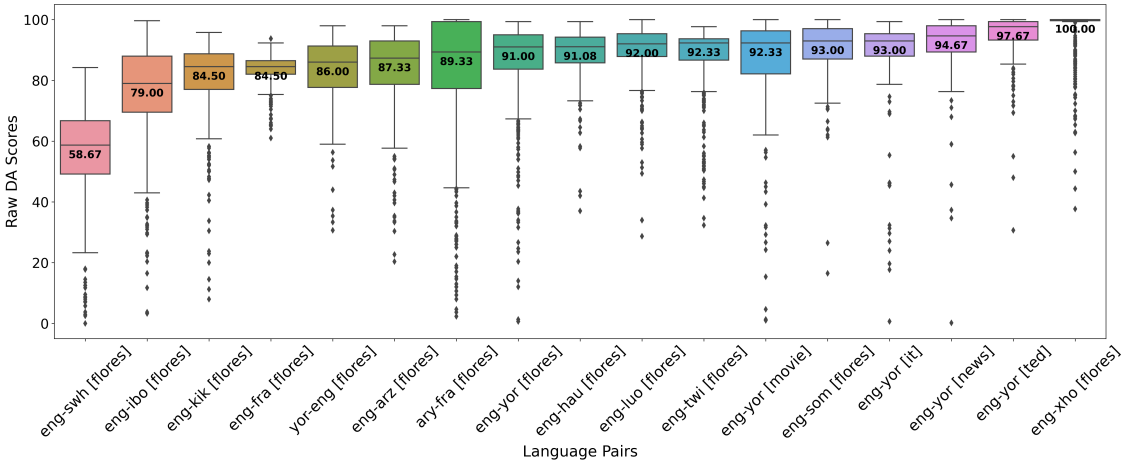

Figure 3: Translation quality of **all** qualified annotated translations as measured by raw DA scores across all language pairs and domains in ascending order, with medians displayed in the plot.

displayed in Figure 2. Each evaluator worked with the tool independently.

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)	250	219	-	219
eng-yor (news)	250	237	-	237
eng-yor (ted)	250	224	-	224
yor-eng	520	439	227	212

Table 1: Counts of qualified annotations for language pairs in dev and devtest sets, with English-Yoruba exclusively as devtest in domain-Specific datasets.

2.3 Quality Assurance

In this study, we implemented a stringent evaluation protocol for each translation result, involving a minimum of **two** bilingual native speakers as evaluators, each with a Bachelor's degree or higher. The evaluators were presented solely with the source text and its corresponding translation. They were encouraged to highlight specific error spans in line with our simplified MQM guidelines first, and then provided a relevant DA score based on the DA guidlines before submission. To prepare for the official annotation process, each evaluator annotated 20 random samples from across all lanugage pairs and domain-specific datasets. Following the individual evaluations, we organized a discussion among the evaluators to review any annotation errors and to address any inconsistencies in their assessments. This preliminary step was designed to familiarize them with the annotation guidelines and the range and nature of translations present in the datasets. It is important to note that, despite strong encouragement, not all evaluators consistently highlighted error spans. As a result, some annotations may have low DA scores but lack corresponding error span highlightings. In our data analysis, particularly when exploring the relationship between detailed issue detection and overall DA scoring, such annotations without error span highlightings will be excluded.

After all evaluators completed their annotations, we compiled the data and excluded translations exhibiting DA score inconsistencies beyond a 34-point range. This threshold, established by our DA guidelines, was critical for ensuring the reliability of our annotation outcomes. To reduce biases among evaluators, we normalized the DA scores at the evaluator level to get the z-scores. The final scores for our benchmark modeling were determined by averaging these z-scores across evaluators for each translation.

After filtering out inconsistencies among evaluators, we present the counts of qualified translation annotations across each language pair within the dev and devtest sets in Table 1: annotations for English-Egyptian Arabic exhibit the highest consis-

tency among evaluators, whereas annotations for English-Igbo show the lowest.

We calculate inter-annotator agreement using the method outlined in (Pavlick and Tetreault, 2016). For each instance, we randomly designate one annotation as Annotator 1, and the average of the other annotations represents the assessment of label 2. Next, we calculate the Pearson correlation coefficient between the two sets of annotations. This procedure of annotation assignment and correlation calculation is repeated 100 times. The resulting inter-annotator agreement for our gathered assessments is 0.797, indicating a strong and consistent agreement among the evaluators.

2.4 Quantitative Analysis

Overall Translation Quality We show the distribution of the raw DA scores across all language pairs and domains in Figure 3. Notably, English-Swahili translations generated by the M2M-100 engine exhibit the lowest translation quality (median DA: 58.67), whereas the English-Xhosa translations produced by the NLLB-200 engine demonstrate the highest translation quality (median DA: 100). Additionally, the Darija-French translations display the greatest variance in translation quality.

Error Counts vs. DA score Equipped with annotation datasets predominantly comprising both overall DA scores and fine-grained error span detections, we aim to investigate the correlation between these two aspects. As previously mentioned in Section 2.3, we encountered instances where annotations had low DA scores without associated error span highlighting. To ensure a focused analysis on the relationship between detailed error spans and overall DA scores, we have selectively filtered out any annotations with DA scores below 80 that do not include error span highlighting. Subsequent to this filtering, we illustrate the counts of error words within each issue category and the associated sentence-level quality scores measured by DA scores in Figure 4. It is evident that Mistranslation emerges as the most common error category across all language pairs, potentially exerting a significant impact on DA scores. Interestingly, English-Yoruba translations in the context of movie-related content exhibit a higher incidence of Omission errors, whereas translations within the IT domain is more prone to Addition errors.

In order to better understand how error categories at the span level influence annotators' judgment at the sentence level, we have calculated and reported Spearman-rank and Kendall-rank correlation coefficients between various issue counts and assessment scores (raw DA scores and normalized z-scores) in Table 2. These coefficients suggest that Mistranslation, as the most prevalent error type, exhibits a moderate to high negative correlation with raw DA scores, indicating its significant influence on the sentence-level DA evaluations of annotators. Furthermore, the total counts of issues correlates strongly and negatively with both raw DA scores and normalized z-scores, further affirming the significance of our adapted simplified MQM guidelines.

CRITERIA	SPEARMAN		KENDALL	
	DA score	Z-score	DA score	Z-score
Mistranslation	-0.675	-0.544	-0.546	-0.422
Omission	-0.318	-0.304	-0.263	-0.246
Addition	-0.207	-0.211	-0.172	-0.172
Untranslated	-0.156	-0.119	-0.130	-0.097
Total Issue	-0.791	-0.687	-0.640	-0.533

Table 2: Correlation between error counts and sentence-level translation quality across various issue categorizes.

3 AFRICOMET: Benchmark System

In this section, we will introduce how we develop our MT evaluation systems for African languages. In our comparative analysis, we benchmark our systems against (1) the widely recognized n-gram and character-based evaluation metrics, SacreBLEU (Post, 2018) and chrf++ (Popović, 2017), and (2) the cutting-edge neural, pre-trained language model-based COMET metric (Rei et al., 2020, 2022a).

3.1 Experimental Settings

3.1.1 Training Data

Beginning in 2017, the organizers of the WMT News translation tasks have been gathering annotations using the Direct Assessment (DA) method (Graham et al., 2013). We employ these DA datasets, which are also utilized by the COMET metric (COMET22) (Rei et al., 2022a), as training data for our systems. In addition, another large sourced DA annotations is the MLQE-PE datasets (Fomicheva et al., 2020), which is typically used for WMT Quality Estimation Shared Tasks (Specia et al., 2020, 2021; Zerva et al., 2022b). The training corpus comprised with DA annotations from the 2017 to 2020 WMT News

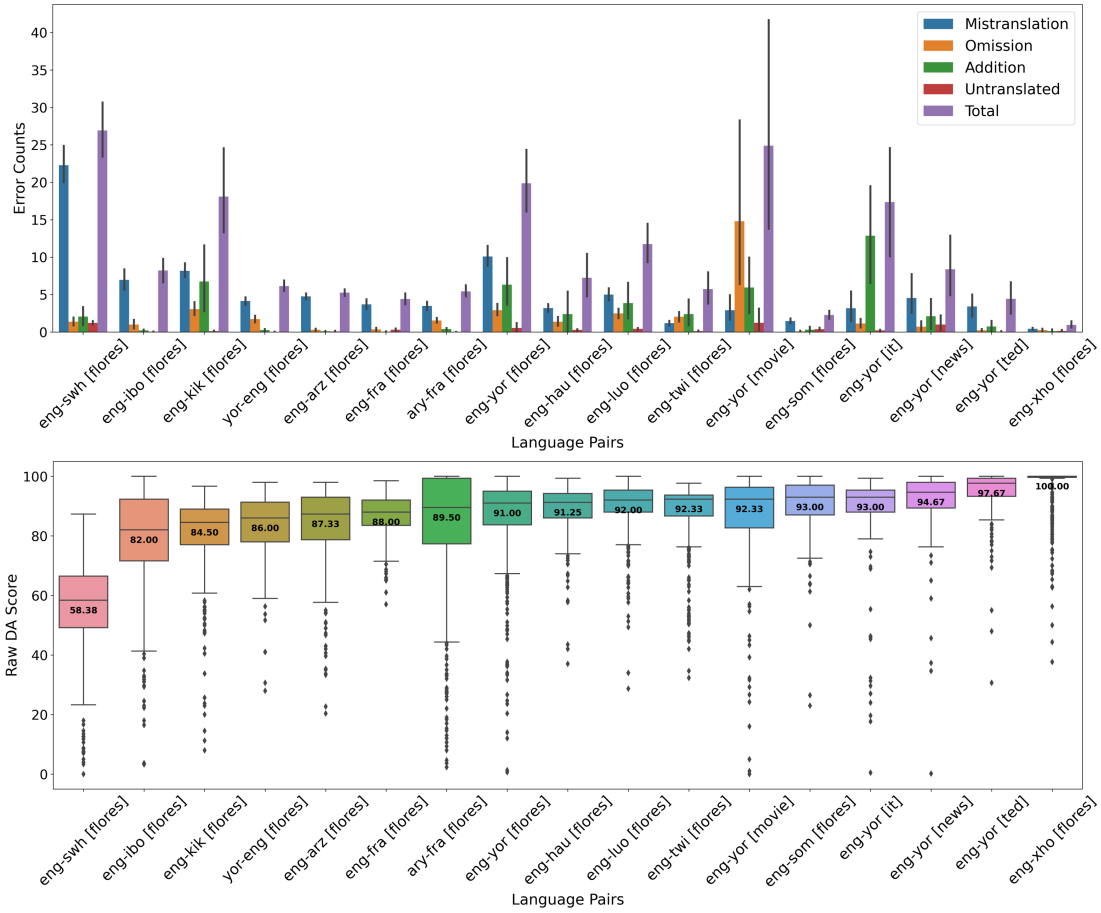

Figure 4: Error Counts with each issue categories and sentence-level translation quality measured by DA scores across all language pairs and domains.

translations and the MLQE-PE dataset is collectively referred to as **"WMT Others"**.

Moreover, a more recent and relevant DA dataset is the human evaluation dataset from the WMT 2022 Large-Scale Machine Translation Shared Task for African Languages (Adelani et al., 2022), comprising human evaluations of 99 source segments from the FLORES-101 test set across eight translation systems for 46 language pairs using the Direct Assessment method. We refer to this annotation dataset as **"WMT African"**.

Statistical summaries of the "WMT Others" and "WMT African" datasets are provided in Table 6 and Table 7, respectively, located in Appendix A. Any duplicate annotations have been excluded from these training sets. During preprocessing, to facilitate interpretability and manage the unbounded nature of the quality scores, we apply the min-max scaling on the normalized z-scores adjusting their range to fall between 0 and 1.

3.1.2 Model configurations

Our investigation revolves 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 potential benefits of an additional MT evaluation dataset in African languages for modeling, and (3) the impact of using African language-enhanced pre-trained models on the performance of MT evaluation.

Our MT evaluation model takes as input the triplets consisting of the source sentence, its machine translation, and the corresponding reference, training a regression model to predict normalized scores that have been scaled between 0 and 1. The dev sets within AFRIMTE will serve as validation sets in the development of our benchmark MT evaluation system for African languages, with the devtest sets designated as test sets.

We construct our models upon the Estimator framework (Rei et al., 2020), utilizing a variety of pre-trained language models as the underlying en-

coders: XLM-Roberta-large (XLM-R-L) (Conneau et al., 2019), InfoXLM-large (InfoXLM-L) (Chi et al., 2020), and two African language enhanced pre-trained models AfriBERTa-large (AfriBERTa-L) (Ogueji et al., 2021) and AfroXLM-Roberta-large (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) for the WMT 2022 MT Evaluation and Quality Estimation Shared Tasks; AfriBERTa-L and AfroXLM-R-L are both based on the XLM-Roberta-large architecture, with the former being a multilingual model trained from scratch on texts in 11 African languages, and the latter adapted the XLM-Roberta model on data from 17 African languages.

We train our models with the open-sourced codebase of COMET metric [6]. Training for 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 of other hyperparameters of the COMET metric.

3.1.3 Evaluation

We report Pearson correlation, Spearman-rank correlation, and Kendall-rank correlation coefficients to assess the correlation between automated scores predicted by our models and the human-annotated scores, with the Spearman-rank correlation coefficient designated as our primary monitoring metric. Additionally, to ensure statistical significance, we user Perm-Both hypothesis test (Deutsch et al., 2021), using 200 re-sampling runs, and setting $p = 0.05$. These evaluations are specifically conducted on the devtest subsets of the AFRIMTE dataset across various language pairs and domains.

3.2 Main findings

3.2.1 Transfer learning

Initially, we develop our MT evaluation systems that leverage transfer learning from a variety of other languages to African languages. Essentially, we train our models on the "WMT Others" dataset and employ the dev sets within our AFRIMTE dataset as validation sets. As outlined in Section 3.1.2, to explore the impact of different pre-trained models on building MT evaluation systems, we conduct experiments based on XLM-R-L, InfoXLM-L, AfriBERTa-L and AfroXLM-R-L for comparison. In addition, we benchmark our models against the COMET metric, specifically COMET22 (Rei et al., 2022a), which uses the same training data but differs in validation, employing additional MQM data for English-German, Chinese-English, and English-Russian from the WMT 2021 News Shared Task.

Results of Spearman-rank correlation coefficients and Perm-Both hypothesis test results are shown in Table 3. Given that "WMT Others" dataset does not include any African language, the results in Table 3 illuminate the effectiveness of different pre-trained models in zero-shot scenarios involving African languages. It is obvious that the AfroXLM-R-L model demonstrates a promising ability to transfer learning from other languages to African languages. It achieves the highest Spearman-rank correlation among various pre-trained models considered on a considerable portion of language pairs.

Analyzing Spearman-rank correlations within four **domain-specific** English-Yoruba datasets show that models trained based on AfroXLM-R-L and InfoXLM-L have the potential to surpass the performance of COMET22. When utilizing the 13 FLORES development sets in African languages as validation sets, these models further illustrate the benefit of targeted pre-trained language models in enhancing domain-specific machine translation evaluation.

Moreover, out of the 17 devtest sets evaluated, AfroXLM-R-L achieves the top rankings in 12 language pairs, surpassing other models except for InfoXLM-L. This underlines the significance of tailored, language-specific pre-trained models in improving downstream NLP performance, especially in linguistically diverse contexts such as those found in African languages.

Another interesting observation is that despite AfriBERTa-L being pre-trained exclusively on African languages, suggesting potential underrepresentation of other languages, it nonetheless proves capable of facilitating transfer learning when trained on languages outside of its initial pre-training scope. However, its overall performance lags notably behind that of AfroXLM-R-L, likely owing to its smaller, Africa-specific pre-training data. To delve deeper into this issue, we conducted experiments using a variety of training data settings, the results of which are detailed in Table 11 in Appendix A. The outcomes of these experiments suggest that exclusive training on African languages

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

	N-gram based Metrics		Baseline	Models based on various Pre-trained Encoders (Ours)			
LP	SacreBLEU	chrf++	COMET22	XLM-R-L	AfroXLM-R-L	AfriBERTa-L	InfoXLM-L
ary-fra	0.332	0.328	0.533	0.551	0.567	0.387	**0.627**
eng-arz	0.324	0.321	0.503	0.486	0.532	0.336	**0.596**
eng-fra	0.246	0.280	**0.489**	**0.510**	**0.495**	**0.446**	**0.525**
eng-hau	0.200	0.301	**0.430**	0.401	**0.515**	0.394	0.378
eng-ibo	0.339	0.424	0.373	0.413	**0.592**	0.453	0.229
eng-kik	0.273	**0.295**	0.202	0.281	**0.389**	**0.298**	**0.303**
eng-luo	0.182	**0.279**	0.062*	**0.201**	**0.283**	**0.239**	**0.232**
eng-som	0.161	0.279	0.474	0.466	**0.554**	0.340	0.412
eng-swh	0.481	0.565	**0.738**	**0.739**	0.688	0.603	**0.773**
eng-twi	**0.204**	**0.178**	0.096*	0.103*	0.157	**0.223**	0.145
eng-xho	0.090*	**0.161**	0.071*	0.070*	**0.191**	**0.151**	0.071*
eng-yor	0.210	0.204	0.150	0.193	**0.287**	**0.270**	**0.313**
eng-yor (it)	0.295	0.346	0.334	0.256	0.266	0.374	**0.487**
eng-yor (movie)	0.238	0.221	**0.334**	**0.338**	**0.372**	**0.325**	**0.353**
eng-yor (news)	**0.114**	**0.122***	**0.168**	**0.196**	**0.200**	0.100	**0.129**
eng-yor (ted)	0.027*	0.002*	0.123*	0.177	**0.324**	0.227	**0.280**
yor-eng	0.308	**0.408**	**0.502**	**0.460**	**0.490**	0.405	**0.461**
Avg. (Spearman / Perm-Both)	0.237 / 2.47	0.277 / 2.06	0.328 / 1.82	0.344 / 1.76	**0.406 / 1.35**	0.328 / 1.88	0.371 / 1.35

Table 3: Spearman-rank correlation coefficients for models trained on the "WMT Others" training set using various pre-trained encoders. Values marked with * indicate a p-value greater than 0.05. For each devtest set, values in **bold** represent the highest ranking obtained from the Perm-Both hypothesis test (Deutsch et al., 2021). Comprehensive results of this test are detailed in Table 8. Averaged Spearman-rank correlation and Perm-Both rankings are presented in the last row.

does not inherently improve model performance and remains inferior to transferring from a broader range of other languages with larger datasets.

3.2.2 Impact of training dataset

To discuss the potential benefits of an additional MT evaluation dataset in African languages, we carry out 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 results, including Pearson, Spearman-rank, Kendall-rank correlation coefficients, and Perm-Both hypothesis test results, are detailed in Table 4.

Surprisingly, the "WMT Others" dataset alone yields highest Spearman-rank and Kendall-rank correlations than the "WMT Combined" dataset. While "WMT Combined" secures the highest Pearson correlation, it slightly negatively impacts both Spearman-rank and Kendall-rank correlations. Examining all of the three correlation coefficients and the Perm-Both hypothesis test results reveals that models trained on "WMT Others" and "WMT Combined" significantly outperform the model trained solely on "WMT African". The difference in performance is likely due to the relatively limited size of "WMT African", indicating a scarcity of data when compared to "WMT Others".

3.3 The benchmark systems

In summary, we have developed two MT evaluation systems that are benchmarked against the state-of-the-art COMET metric, achieving a Spearman-rank correlation with human judgments of up to +0.406 for African languages. These systems are trained using the "WMT Others" and "WMT Combined" datasets, with the Afro-XLM-Roberta-large model serving as the pre-trained foundation for model training.

4 Reference-free Evaluation

Utilizing the annotated AFRIMTE dataset, we are able to develop reference-free models that predict the quality of machine translations in the absence of reference texts. This approach aligns with the research domain of machine translation quality estimation (QE), as explored in works by Specia et al. (2010), Fan et al. (2019), Kepler et al. (2019), Chatzikoumi (2020), and Ranasinghe et al. (2020). For this purpose, we adopt the architecture of our MT evaluation models, albeit without including references in the input. Selecting AfroXLM-R-L and InfoXLM-L as pre-trained models, we train reference-free models under the "WMT Others" and "WMT Combined" data settings. We compare our developed reference-free QE systems with CometKiwi (Rei et al., 2022b), which also adopts

	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. (Corr / Perm-Both)	0.481 / 1.94	0.261 / 1.94	0.179 / 2.12	0.531 / 1.65	**0.406 / 1.00**	**0.288 / 1.12**	**0.594 / 1.06**	0.393 / 1.12	0.279 / 1.18

Table 4: Correlation coefficients (Pearson, Spearman-rank, Kendall-rank) for models trained based on AfroXLM-Roberta-Large with varied training data settings. Values marked with * indicate a p-value greater than 0.05. For each devtest set, values in **bold** represent the highest ranking obtained from the Perm-Both hypothesis test (Deutsch et al., 2021). Comprehensive results of this test are detailed in Table 9. The average of correlation coefficient (Corr) and Perm-Both rankings are presented in the last row.

	Baseline		**Models based on Various Pre-trained Encoders (Ours)**							
	CometKiwi		InfoXLM-L				AfroXLM-R-L			
			WMT Others		WMT Combined		WMT Others		WMT Combined	
LP	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
ary-fra	0.517	**0.495**	**0.578**	**0.526**	**0.591**	**0.515**	0.475	**0.507**	0.398	**0.459**
eng-arz	**0.611**	0.592	**0.612**	**0.616**	0.593	0.599	0.551	0.516	0.538	0.509
eng-fra	**0.527**	0.495	**0.552**	**0.522**	**0.553**	**0.523**	0.418	0.478	0.391	0.399
eng-hau	0.314	0.245	0.364	0.241	**0.737**	0.275	0.652	**0.482**	0.670	0.435
eng-ibo	0.205	0.188	0.175	0.181	0.287	0.200	**0.644**	**0.631**	**0.628**	**0.631**
eng-kik	0.277	0.247	0.322	0.283	**0.642**	0.307	0.631	**0.415**	**0.687**	**0.449**
eng-luo	0.237	0.161	**0.253**	0.171	**0.397**	**0.190**	**0.333**	**0.217**	**0.344**	**0.211**
eng-som	0.266	0.357	0.281	0.327	**0.394**	0.353	**0.302**	**0.482**	0.260	0.442
eng-swh	**0.787**	**0.756**	**0.803**	**0.763**	**0.790**	**0.750**	0.644	0.587	0.596	0.541
eng-twi	0.097*	0.026*	0.135	0.061*	**0.546**	**0.136**	0.290	0.061	**0.390**	0.058*
eng-xho	0.127	-0.030*	0.151	0.006*	**0.543**	**0.071***	**0.437**	**0.085***	**0.447**	**0.085***
eng-yor	0.327	0.231	0.354	0.280	**0.767**	0.232	0.738	**0.392**	**0.763**	**0.437**
eng-yor (it)	0.375	**0.388**	0.385	**0.402**	**0.822**	**0.328**	0.654	**0.318**	0.730	**0.311**
eng-yor (movie)	0.151	0.041*	0.215	0.087*	**0.710**	**0.336**	0.557	**0.314**	0.611	**0.353**
eng-yor (news)	0.104*	0.078*	0.130	**0.088***	**0.563**	-0.035*	**0.508**	**0.186**	**0.529**	**0.208**
eng-yor (ted)	0.217	**0.289**	0.223	**0.274**	**0.480**	0.082*	**0.518**	**0.189**	**0.535**	**0.202**
yor-eng	0.070*	0.098*	0.097	0.122*	**0.342**	**0.265**	0.181	0.208	**0.284**	**0.295**
Avg. (Corr / Perm-Both)	0.306 / 2.65	0.274 / 2.12	0.331 / 2.41	0.291 / 1.76	**0.574 / 1.18**	0.302 / 1.59	0.502 / 1.76	**0.357 / 1.35**	0.518 / 1.59	0.354 / 1.53

Table 5: Correlation coefficients (Pearson, Spearman-rank) for reference-free QE models trained based on AfroXLM-Roberta-Large and InfoXLM-Large with varied training data settings. Values marked with * indicate a p-value greater than 0.05. For each devtest set, values in **bold** represent the highest ranking obtained from the Perm-Both hypothesis test (Deutsch et al., 2021). Comprehensive results of this test are detailed in Table 10. The average of correlation coefficient (Corr) and Perm-Both rankings are presented in the last row.

"WMT Others" for training data and is built based on the InfoXLM-L architecture.

QE systems are typically evaluated based on Pearson and Spearman-rank correlations, as highlighted by Zerva et al. (2022b), and our experimental results are presented in Table 5. The results demonstrate that the model trained on the "WMT Others" dataset substantially surpass the baseline CometKiwi system under the same pre-trained model setting, namely InfoXLM-L in terms of both Pearson and Spearman-rank correlations. This underscores the viability of applying transfer learning from DA datasets in other languages to African languages for the MT quality estimation task. Additionally, incorporating DA datasets in African languages notably enhances the Pearson correlation across both training data settings, with a more pronounced improvement using InfoXLM-L and a slight enhancement with AfroXLM-R-L. Moreover, employing African language-enhanced

pre-trained models further boosts performance in both Pearson and Spearman-rank correlations.

Finally, a key area of our research involves examining the disparity between the more challenging reference-free models and the simpler reference-based models in MT quality estimation, to deepen our understanding and bridge this gap. As illustrated in Tables 3 and 4, there is a Spearman-rank correlation gap of 0.04 $(0.371 - 0.331)$ when using InfoXLM-L as the pre-trained model and training on the 'WMT Others' dataset, in comparison to reference-based models. Additionally, when utilizing AfroXLM-R-L as the pre-trained model, the gaps in Pearson correlation are 0.029 $(0.531 - 0.502)$ and 0.076 $(0.594 - 0.518)$ for the "WMT Others" and "WMT Combined" training datasets, respectively. Similarly, the gaps in Spearman-rank correlation are 0.049 $(0.406 - 0.357)$ and also 0.049 $(0.393 - 0.354)$ for training on "WMT Others" and "WMT Combined", respectively. All these gaps in correlation coefficients are less than 0.05, indicating a relatively close performance between the MT evaluation models and the MT quality estimation models, highlighting key areas for further development in efficient and accurate MT quality estimation models. Overall, both InfoXLM-L and AfroXLM-R-L are promising in building superior QE systems compared to the state-of-the-art CometKiwi system.

5 Conclusion

This study tackles the challenges of adapting the COMET metric for machine translation evaluation in various under-resourced African languages. We have developed a simplified MQM annotation guideline, created the AFRIMTE dataset encompassing 13 typologically-diverse African languages, and established benchmark COMET systems AFRICOMET, thereby addressing pivotal issues in this domain. Our dedication to open access, demonstrated by the release of all datasets, code, and models, aims to bolster ongoing research and development in the field of machine translation evaluation.

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.

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.

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, 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. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, 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.

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.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André FT Martins. 2019. Openkiwi: An open source framework for quality estimation. *arXiv preprint arXiv:1902.08646*.

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.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126.

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, Copenhagen, 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.

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

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.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine translation*, 24:39–50.

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 Example Appendix

This is an appendix.

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 Count	1027155			

Table 6: 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	30022			

Table 7: 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).

	N-gram based Metrics		Baseline	Models based on various Pre-trained Encoders (Ours)			
LP	SacreBLEU	chrf++	COMET22	XLM-R-L	AfroXLM-R-L	AfriBERTa-L	InfoXLM-L
ary-fra	3	3	2	2	2	3	1
eng-arz	4	4	2	3	2	4	1
eng-fra	2	2	1	1	1	1	1
eng-hau	4	3	1	2	1	2	2
eng-ibo	2	2	2	2	1	2	3
eng-kik	2	1	3	2	1	1	1
eng-luo	2	1	3	1	1	1	1
eng-som	4	3	2	2	1	3	2
eng-swh	4	3	1	1	2	3	1
eng-twi	1	1	2	2	2	1	2
eng-xho	2	1	2	2	1	1	2
eng-yor	2	2	3	2	1	1	1
eng-yor (it)	2	2	2	3	3	2	1
eng-yor (movie)	2	2	1	1	1	1	1
eng-yor (news)	1	1	1	1	1	2	1
eng-yor (ted)	3	3	2	2	1	2	1
yor-eng	2	1	1	1	1	2	1
Average	2.47	2.06	1.82	1.76	1.35	1.88	1.35

Table 8: Detailed rankings from the Perm-Both hypothesis test (Deutsch et al., 2021) of Spearman-rank correlation coefficients for models trained on the "WMT Others" training set using various pre-trained encoders. The averaged ranks are presented in the last row.

	Training Data Settings								
	Pearson			Spearman			Kendall		
LP	WMT African	WMT Others	WMT Combined	WMT African	WMT Others	WMT Combined	WMT African	WMT Others	WMT Combined
ary-fra	3	1	1	2	1	1	2	1	1
eng-arz	3	1	1	3	1	2	3	1	2
eng-fra	2	1	1	2	1	1	2	1	1
eng-hau	2	2	1	2	1	1	3	1	1
eng-ibo	3	1	2	2	1	1	2	1	1
eng-kik	1	2	1	2	1	1	2	1	1
eng-luo	1	1	1	1	1	1	1	1	1
eng-som	2	2	1	2	1	1	2	1	1
eng-swh	2	1	1	2	1	1	3	2	2
eng-twi	1	2	1	1	1	1	2	2	1
eng-xho	1	2	1	2	1	1	2	1	1
eng-yor	2	3	1	2	1	1	2	1	1
eng-yor (it)	2	3	1	1	1	1	1	1	1
eng-yor (movie)	2	3	1	2	1	1	2	1	1
eng-yor (news)	2	1	1	2	1	1	2	1	1
eng-yor (ted)	2	1	1	3	1	2	3	1	2
yor-eng	2	1	1	2	1	1	2	1	1
Average	1.94	1.65	1.06	1.94	1	1.12	2.12	1.12	1.18

Table 9: Detailed rankings from the Perm-Both hypothesis test (Deutsch et al., 2021) of Pearson, Spearman-rank, Kendall-rank correlation coefficients for models trained based on AfroXLM-Roberta-Large with varied training data settings. The averaged ranks are presented in the last row.

Correlation	LP	Baseline	InfoXLM-L		AfroXLM-R-L	
		CometKiwi	WMT Others	WMT Combined	WMT Others	WMT Combined
Pearson	ary-fra	2	1	1	2	3
	eng-arz	1	1	2	2	2
	eng-fra	1	1	1	2	2
	eng-hau	4	3	1	2	2
	eng-ibo	3	4	3	1	1
	eng-kik	4	3	1	2	1
	eng-luo	2	1	1	1	1
	eng-som	2	2	1	1	2
	eng-swh	1	1	1	2	3
	eng-twi	3	3	1	2	1
	eng-xho	2	2	1	1	1
	eng-yor	3	3	1	2	1
	eng-yor (it)	4	4	1	3	2
	eng-yor (movie)	5	4	1	3	2
	eng-yor (news)	3	3	1	1	1
	eng-yor (ted)	2	2	1	1	1
	yor-eng	3	3	1	2	1
	Average	2.65	2.41	1.18	1.76	1.59
Spearman	ary-fra	1	1	1	1	1
	eng-arz	2	1	2	3	3
	eng-fra	2	1	1	2	3
	eng-hau	3	3	3	1	2
	eng-ibo	3	3	3	1	1
	eng-kik	2	2	2	1	1
	eng-luo	2	2	1	1	1
	eng-som	2	3	2	1	2
	eng-swh	1	1	1	2	3
	eng-twi	2	2	1	2	2
	eng-xho	3	2	1	1	1
	eng-yor	3	2	2	1	1
	eng-yor (it)	1	1	1	1	1
	eng-yor (movie)	3	2	1	1	1
	eng-yor (news)	2	1	2	1	1
	eng-yor (ted)	1	1	2	1	1
	yor-eng	3	2	1	2	1
	Average	2.12	1.76	1.59	1.35	1.53

Table 10: Detailed rankings from the Perm-Both hypothesis test (Deutsch et al., 2021) of Pearson and Spearman-rank correlation coefficients for reference-free QE models trained based on AfroXLM-Roberta-Large and InfoXLM-Large with varied training data settings. 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	0.255	0.214	0.154	0.409	0.387	0.275	0.385	0.384	0.275
eng-arz	0.091*	0.111*	0.075*	0.305	0.336	0.226	0.307	0.312	0.205
eng-fra	0.341	0.292	0.202	0.430	0.446	0.315	0.418	0.429	0.301
eng-hau	0.608	0.328	0.224	0.573	0.394	0.276	0.722	0.399	0.276
eng-ibo	0.292	0.300	0.211	0.379	0.453	0.317	0.340	0.475	0.327
eng-kik	0.559	0.229	0.155	0.463	0.298	0.207	0.477	0.231	0.160
eng-luo	0.285	0.146	0.099	0.364	0.239	0.163	0.378	0.189	0.127
eng-som	0.361	0.256	0.180	0.332	0.340	0.240	0.380	0.302	0.212
eng-swh	0.522	0.464	0.323	0.665	0.603	0.434	0.660	0.558	0.394
eng-twi	0.411	0.223	0.150	0.367	0.223	0.150	0.381	0.224	0.150
eng-xho	0.391	0.104*	0.082*	0.370	0.151	0.119	0.461	0.122	0.096
eng-yor	0.564	0.180	0.122	0.388	0.270	0.186	0.615	0.218	0.148
eng-yor (it)	0.721	0.430	0.303	0.595	0.374	0.257	0.766	0.300	0.207
eng-yor (movie)	0.511	0.320	0.217	0.447	0.325	0.227	0.588	0.364	0.252
eng-yor (news)	0.339	0.043*	0.029*	0.313	0.100	0.066	0.488	0.105	0.068
eng-yor (ted)	0.403	0.159	0.109	0.357	0.227	0.154	0.534	0.166	0.114
yor-eng	0.215	0.218	0.144	0.397	0.405	0.279	0.377	0.390	0.267
Avg. (Corr)	0.404	0.236	0.163	0.421	**0.328**	**0.229**	**0.487**	0.304	0.211

Table 11: Correlation coefficients (Pearson, Spearman-rank, Kendall-rank) for models trained based on AfriBERTa-L with varied training data settings. Values marked with * indicate a p-value greater than 0.05. The average of correlation coefficient (Corr) are presented in the last row.