SSA-COMET: Do LLMs Outperform Learned Metrics in Evaluating MT for Under-Resourced African Languages?

Anonymous ACL submission

Abstract

Evaluating machine translation (MT) quality for under-resourced African languages remains a significant challenge, as existing metrics often suffer from limited language coverage and poor performance in low-resource settings. While recent efforts, such as AfriCOMET, have addressed some of the issues, they are still constrained by small evaluation sets, a lack of publicly available training data tailored to African languages, and inconsistent performance in extremely low-resource scenarios. In this work, we introduce ***SSA-MTE***, a large-scale human-annotated MT evaluation (MTE) dataset covering 13 African language pairs from the News domain, with over 63,000 sentence-level annotations from a diverse set of MT systems. Based on this data, we develop ***SSA-COMET*** and ***SSA-COMET-QE***, improved reference-based and reference-free evaluation metrics. We also benchmark prompting-based approaches using state-of-the-art LLMs like *GPT-4o*, *Claude-3.7* and *Gemini 2.5 Pro* . Our experimental results show that SSA-COMET models significantly outperform AfriCOMET and are competitive with the strongest LLM (*Gemini 2.5 Pro*) evaluated in our study, particularly on low-resource languages such as Twi, Luo, and Yorùbá. All resources are released under open licenses to support future research.

1 Introduction

Recent advancements in machine translation evaluation (MTE) have largely benefited high-resource languages. Neural metrics such as COMET and MetricX (Rei et al., 2020; Juraska et al., 2023) have demonstrated strong performance by capturing deeper semantic relationships in translations. However, their effectiveness diminishes for under-represented languages, such as many African languages, due to the scarcity of high-quality training and evaluation data, as well as the limitations in the multilingual large language models used as their pretrained backbones (Freitag et al., 2024; Sai B et al., 2023; Wang et al., 2024b).

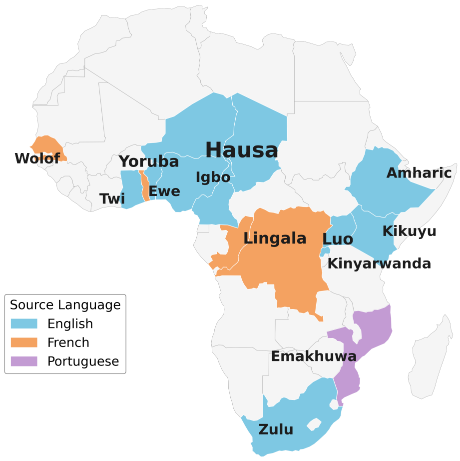

Figure 1: Language distribution across the 13 Sub-Saharan African languages in SSA-MTE.

To narrow this gap, Wang et al. (2024a) introduced AfriMTE, a high-quality evaluation dataset covering 13 typologically diverse African languages, annotated using a simplified version of the MQM framework (Lommel et al., 2014), specifically designed for non-expert annotators. Building on AfriMTE, they developed AfriCOMET, an enhanced version of COMET (Rei et al., 2020), by incorporating an African-centric encoder, AfroXLM-R (Alabi et al., 2022). More recently, Wang et al. (2024b) enhanced these models by adopting AfroXLMR-76, which covered more African languages (Adelani et al., 2024).

However, despite the advances of Wang et al. (2024a,b), several limitations remain. First, the lack of training data in AfriMTE restricts opportunities for the broader research community to improve upon existing models. Second, the evaluation setup of AfriMTE includes only a single MT system per language pair, limiting the diversity of translation outputs and making it challenging to assess the metric's generalizability across systems of varying quality and style. Third, the evaluation datasets in AfriMTE are relatively small—typically around 100–200 annotated examples per language pair—which may not adequately capture the full range of linguistic variation. Finally, AfriCOMET models exhibit unreliable performance for certain extreme low-resource African languages, such as Twi and Luo, producing inconsistent or low-quality estimates (Adelani et al., 2025).

In this work, we address these challenges through three key contributions: **(1)** We expand the landscape of high-quality MT training and evaluation data by introducing ***SSA-MTE***, a new human-annotated dataset covering 13 Sub-Saharan African language pairs, 7 of these pairs are newly introduced compared to AfriMTE. Our annotations are sourced from the *News domain*—selected for its topical diversity, timeliness, and widespread use in the MT community. The machine translated outputs in SSA-MTE are generated using a diverse set of MT systems such as Google Translate and NLLB (NLLB-Team et al., 2022), and frontier large language models (LLMs) such as *GPT-4o* and *Gemini*. **(2)** We enhance the AfriCOMET models by extending them on our newly collected data, resulting in ***SSA-COMET*** and ***SSA-COMET-QE***, improved MTE and reference-free quality estimation (QE) metrics specifically tailored to African languages. **(3)** We fully explore the capabilities of cutting-edge LLMs, including *Gemini 2.5 Pro*, *GPT-4o*, and *Claude 3.7*, for MTE and QE in a few-shot setting on the testing data of SSA-MTE.

Our experimental results demonstrate substantial overall performance improvements of the SSA-COMET models over AFRICOMET-v1.1 (Wang et al., 2024b), with particularly strong gains on low-resource languages such as Twi, Luo, and Yorùbá. In MT evaluation, SSA-COMET demonstrates competitive performance with *Gemini 2.5 Pro* and outperforms other prompting-based LLM metrics, achieving higher average Spearman correlation than *GPT-4o* and *Claude-3.7*, despite being an order of magnitude smaller in model size. To support future research in African NLP and foster reproducibility, we release our dataset, models, and training pipeline under open licenses.

2 Related Works

Traditional MTE metrics like BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ChrF (Popović, 2015) rely on n-gram overlap and correlate poorly with human judgments. Neural metrics such as BERTScore (Zhang et al., 2020) better capture semantic similarity. COMET (Rei et al., 2020) improves on this by framing MTE as a regression task using XLM-R (Conneau et al., 2019) and training data of quality scores. Its extension, COMETKiwi (Rei et al., 2022), removes the need for reference translations, increasing flexibility. More recently, MetricX (Juraska et al., 2023), which is built on mT5 (Xue et al., 2020), adopts a regression-based framework similar to COMET. In parallel, with the rise of LLMs, there is growing interest in prompting LLMs directly to assess translation quality (Kocmi and Federmann, 2023; Freitag et al., 2024).

Recent studies (Wang et al., 2024a,b; Freitag et al., 2024) show that both neural metrics and prompting-based methods perform poorly on under-represented African languages, when compared to high-resource settings. To address this, AfriCOMET (Wang et al., 2024a) uses an Africa-centric encoder, AfroXLMR (Alabi et al., 2022), and Non-African MTE training data to build a COMET-style metric, showing robust performance on African MTE tasks. However, recent analysis (Adelani et al., 2025) finds that AfriCOMET still shows inconsistencies with human judgments in extreme low-resource languages like Twi.

In this paper, we expand the landscape of high-quality MT training and evaluation data for African languages by introducing a newly annotated MTE dataset, and evaluate performance on newly trained COMET-based models and LLMs.

3 SSA-MTE: The Dataset

This section describes the source data and MT systems used to construct SSA-MTE, presents the annotation guidelines and procedure, outlines the quality assurance measures, and provides a quantitative analysis of the resulting dataset.

3.1 Source Data Collection

The News Domain Given the rich structure and high quality of content in the News domain, this work focuses on the News domain, unlike AfriMTE (Wang et al., 2024a), which centers on Wikipedia data. We sourced the input sentences from the news platform *Global Voices*[1], which publishes articles in parallel across multiple languages. Each article is tagged with topical categories such

[1] https://globalvoices.org/

as *Economics & Business* and *Education* to indicate its thematic focus. Translations on Global Voices are produced manually by a global network of volunteer contributors as part of its Lingua program, and all content is published under a Creative Commons Attribution 3.0 (CC BY 3.0) license.

The Source Data Considering that the two dominant official languages in Africa are English and French, we selected all articles available in both languages, totaling $20,419$. From this pool, we filtered for articles tagged with African regions—such as "Guinea-Bissau" and "Gambia"—to ensure the content was relevant to Africa. To avoid potentially sensitive topics, we heuristically excluded articles tagged with categories such as "war-conflict", resulting in a subset of $3,681$ articles. Finally, we used Gemini to automatically detect and remove any remaining content that might be harmful, yielding a final collection of $1,901$ articles. From this refined set, we manually selected 200 articles by reviewing their titles and tags to ensure diverse topical coverage. At the document level, articles were segmented into sentences using the NLTK sentence tokenizer[2]. We then applied fasttext language identification (Joulin et al., 2016) and sentence alignment using LASER (Artetxe and Schwenk, 2019). Sentences were retained if the language confidence score exceeded 99%, and sentence pairs were aligned if their similarity score was above 92.5%. After final deduplication, we obtained $1,500$ distinct parallel English–French sentence pairs for our source sentences.

Choice of the Language Pairs (LP) Given the English–French language pair, we decided to expand the coverage to 12 typologically diverse *Sub-Saharan African languages*—9 using English, and 3 using French as the source language, to reflect both the Anglophone and Francophone linguistic diversity in the region. We excluded North African languages, as the most widely spoken languages in the region are Arabic dialects, which tend to yield reliable evaluation results with existing metrics such as COMET (Wang et al., 2024a). The English–sourced pairs include Amharic (`eng-amh`), Hausa (`eng-hau`), Igbo (`eng-ibo`), Kikuyu (`eng-kik`), Kinyarwanda (`eng-kin`), Luo (`eng-luo`), Twi (`egn-twi`), Yorùbá (`eng-yor`), and Zulu (`eng-zul`); while the French–sourced pairs include Ewe (`fra-ewe`), Lingala (`fra-lin`), and Wolof (`fra-wol`). Additionally, we include one extreme low-resource Mozambique language, Emakhuwa (`vmw`), sourced from Portuguese (`por`) as detailed in Appendix C.

[2] `https://www.nltk.org/api/nltk.tokenize.html`

3.2 MT Systems

To ensure a diverse representation of translation quality and styles, we used six MT systems to generate translation outputs: four closed-source models including *GPT-4o*, *Gemini-1.5*, *Claude-3.5* [3], and *Google Translate*, and two open-source models including *NLLB-200-distilled-600M* (NLLB-Team et al., 2022) and *M2M-100-418M* (Fan et al., 2021). Since M2M-100 does not support certain languages such as Ewe and Kikuyu, we fine-tuned a separate model for each of these languages using $500,000$ randomly selected samples from the NLLB dataset[4] to ensure consistent translation quality. During this procedure, Kikuyu was not supported by Google Translate; therefore, translations for this language were generated using only five systems. Similarly, for Ewe and Wolof, we excluded GPT-4o outputs, as the model declined to produce translations in more than half of the cases. The MT outputs for `por-vmw` are detailed in Appendix D.

3.3 Annotation Guidelines, Tool, and Protocol

Building on the success of the simplified MQM annotation guidelines proposed by Wang et al. (2024a), we adopt the same framework for both error-span and scoring annotations in this work. Specifically, we evaluate the adequacy of each machine translation output. Evaluators review both the source and translated texts, highlighting error spans, categorized as "`Addition`", "`Omission`", "`Mistranslation`", and "`Untranslated`". They then assign an overall translation quality score using a continuous direct assessment (DA) scale ranging from 0 to 100, strictly following the annotation protocol established in Wang et al. (2024a).

We used the same annotation tool introduced in Wang et al. (2024a),[5] which provides an interface supporting both error span highlighting and DA scoring, and allows each evaluator to work independently. For each LP, we recruited *two bilingual native speakers* with at least a Bachelor's degree to serve as evaluators. Annotation work was

[3] A template for prompting LLMs for translations is provided in Figure 5.

[4] `https://huggingface.co/datasets/allenai/nllb`

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

evenly divided, with 300 overlapping samples included to assess inter-evaluator agreement for quality assurance. Reference translations per LP were produced by two professional translators, who manually translated the sources from scratch, without using any machine translation tools. We annotated 6,600 samples per language pair, including 300 overlapping samples for inter-evaluator agreement, and this results in *6,300 distinct samples* per LP, evenly distributed across MT systems.[6] For each LP, all 1,500 source sentences were translated into the target African language.

3.4 Annotation Quality Assurance

We employed several measures to assure the quality of the annotated data.

Evaluator Selection To select qualified evaluators from a candidates pool, we followed the training procedure outlined in Wang et al. (2024a). Each candidate was required to complete an annotation test designed to both familiarize them with the annotation tool and evaluate their understanding of the annotation guidelines. The test included 22 samples: 20 unique samples drawn from the dataset and 2 repeated samples to assess self-consistency. We assessed the submitted annotations using a heuristic quality check. Specifically, we flagged cases where the assigned score and the highlighted error spans were inconsistent—for example, when a score below 80 was assigned without any error spans, or when a score of 100 was given despite the presence of errors. Moreover, Inter-evaluator agreement was measured by checking whether score differences were below 20 among evaluators. For the repeated samples, we evaluated each candidate's self-consistency, defined as producing similar error spans and assigning scores that differed by less than 5. Finally, a manual review was conducted to ensure overall annotation quality. For each LP, we select the top two evaluators who satisfied four criteria: (1) more than 80% agreement with each other, (2) minimal heuristic quality issues, (3) high self-consistency, and (4) a satisfactory outcome in manual quality review.

Agreement on the Overlaps After selecting the evaluators, we implemented a quality assurance procedure using annotations on the 300 overlapping samples. These samples were independently annotated by both evaluators and served to assess inter-evaluator agreement. To evaluate annotation quality and consistency, we computed Spearman-rank and Pearson correlation coefficients, as well as the Intraclass Correlation Coefficient (ICC) between their assigned scores. Since the evaluators were fixed for each language pair (i.e., the only raters of interest), we used the two-way mixed-effects model ICC(3,,k), with $k = 2$ in our setup.

LP	Pearson	Spearman	ICC(3,2)
eng-amh	0.597	0.653	0.747
eng-hau	0.406	0.476	0.573
eng-ibo	0.314	0.253	0.358
eng-kik	0.735	0.776	0.847
eng-kin	0.486	0.513	0.632
eng-luo	0.735	0.724	0.842
eng-twi	0.757	0.772	0.862
eng-yor	0.567	0.520	0.723
eng-zul	0.249	0.107	0.392
fra-ewe	0.560	0.612	0.694
fra-lin	0.399	0.339	0.570
fra-wol	0.592	0.648	0.741
por-vma	0.620	0.580	0.764

Table 1: Inter-annotator agreement metrics (Pearson, Spearman-rank, ICC(3,2) on 300 overlapping samples.

To reduce evaluator bias, we first normalized the DA scores at the evaluator level, converting them to z-scores. We then computed the agreement statistics described above on the *300 overlapping samples*. The results are presented in Table 1. LPs that exhibited at least a moderate level of agreement, defined as having *both Spearman rank and Pearson correlation coefficients above* 0.4 *and an ICC value above* 0.5—were included in the training, development, and test sets. As a result, 9 LPs were selected for inclusion in all three data splits: `eng-amh`, `eng-hau`, `eng-kik`, `eng-kin`, `eng-luo`, `eng-twi`, `eng-yor`, `fra-ewe`, and `fra-wol`[7]. Although the remaining LPs did not meet the criteria, we retained them for training to introduce additional language diversity, which may help improve the robustness and generalization for modeling.

Among the three remaining LPs (`eng-ibo`, `eng-zul`, and `fra-lin`), `fra-lin` showed a Pearson correlation close to 0.4 and an ICC above 0.5, indicating moderate positive correlation and agreement, though its Spearman rank correlation was slightly lower at 0.339. Given its relatively acceptable agreement levels, we included `fra-lin` in the training data without additional filtering. In contrast, for `eng-ibo` and `eng-zul`, which exhibited weaker agreement across all metrics, we applied further filtering to remove low-quality annotations

[6]For languages not supported by certain MT systems, annotations were distributed across five systems instead of six.

[7]For `por-vmw`, we only annotated the test data.

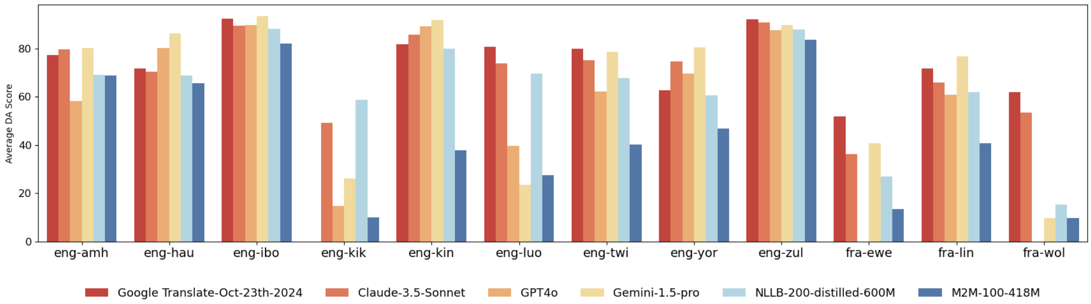

Figure 2: **Average DA scores across MT systems and LPs**. Low-resource pairs such as `eng-kik` and `fra-wol` remain particularly challenging for current translation systems.

LP	Train	Dev	Test
`eng-amh`	4563	326	1166
`eng-hau`	4693	338	1192
`eng-ibo`	1501	–	–
`eng-kik`	4752	318	1172
`eng-kin`	4768	349	1210
`eng-luo`	4691	341	1199
`eng-twi`	4820	325	1200
`eng-yor`	4717	333	1206
`eng-zul`	1905	–	–
`fra-ewe`	4423	296	1077
`fra-lin`	4626	–	–
`fra-wol`	4874	341	1175
`por-vma`	–	–	930
Total	**50333**	**2967**	**11527**

Table 2: Number of **training**, **development**, and **test** examples in SSA-MTE for each LP.

before including them in training. The detailed filtering process is described in Appendix F.

3.5 Additional Test Set in Emakhuwa

To introduce further diversity in evaluation, we include a previously under-studied LP—***Portuguese to Emakhuwa (vmw)***, from Mozambique, in our test set. The data follows the same design as the other 12 LPs: focusing on the *news domain*, *multiple MT systems* are included, and the same *annotation and quality assurance procedures* are applied. Details of the data collection and MT generation processes are provided in Appendix C, D, and E.

3.6 Final Data Statistics

For the final version of the dataset, we applied several filtering steps to ensure high-quality annotations. First, we *excluded all cases with a score below 80* that lacked annotated error spans. We also removed cases falling in the top 20% of DA scores but within the bottom 20% of ChrF scores relative to the reference translations. Similarly, we filtered out cases in the bottom 20% of DA scores that had the highest 20% of ChrF scores.

To avoid potential information leakage, **DEV** and **TEST** sets were selected based on source documents: we *excluded the 300 overlapping examples* used for inter-evaluator agreement and randomly sampled *40 source documents* for the TEST set and *10 documents* for the DEV set. For all languages, only translations whose source sentences came from these selected documents were included in the DEV and TEST sets. This document-level selection helps prevent models from learning translation patterns from highly similar source texts. The *remaining data* was assigned to the **TRAIN** set. Final dataset statistics are reported in Table 2.

To view the translation quality of each MT system for each LP, we present the average DA scores across LPs and MT systems in Figure 2. High-resource LPs, such as `English–Zulu`, generally achieve higher scores, whereas low-resource pairs like `English–Kikuyu` and `French–Wolof` exhibit substantially lower translation quality.

4 SSA-COMET Models

In this section, we describe the modeling approaches of SSA-COMET and SSA-COMET-QE.

4.1 Modeling Methods

MTE Modeling We follow the modeling setup the same as **COMET** for developing MTE systems for African languages. Our models are trained to predict DA adequacy scores, using the **COMET** architecture, which is based on a regression-based estimator framework. We implement both single-task learning (STL) and multi-task learning (MTL).

Single-Task Learning (STL) In the STL setting, each of the source (`src`), machine translation (`mt`), and reference (`ref`) segments is independently encoded using a multilingual encoder. The resulting sentence embeddings are pooled, concatenated, and passed through a feed-forward regressor trained to minimize mean squared error against the human-annotated adequacy scores.

Multi-Task Learning (MTL) In the MTL setting, we adopt the unified multi-view formulation from Wan et al., 2022, where the model is trained **jointly** on three input configurations: ⟨`src, mt`⟩, ⟨`mt, ref`⟩, and ⟨`src, mt, ref`⟩. Each configuration is passed through the model to produce a separate prediction, and the final score is computed by averaging the three outputs. This formulation leverages multiple input perspectives to provide richer supervision and improve generalization.

QE Modeling Additionally, we develop SSA-COMET-QE, a variant that mirrors the AfriCOMET-QE architecture. This model operates solely on the ⟨`src, mt`⟩ pair and is optimized for the QE setting. It is trained independently using the same DA scores, enabling direct quality estimation without relying on reference translations.

5 Experiment Setup

For the TEST evaluation, to ensure comparability across language pairs and annotators, all human-annotated DA scores in the test set were first standardized using z-score normalization.

SSA-COMET training We combine the training data used for AfriCOMET (the WMT Non-African DA data) with the training split of the newly annotated SSA-MTE. Score pre-processing is conducted in two steps: we first apply z-normalization at the evaluator level, followed by min-max scaling to improve consistency and interpretability. To establish a stable global range, we collect the 800 highest and 800 lowest z-scores across all languages and use their corresponding averages to define the minimum and maximum values. The resulting scores are then scaled and clipped to fall within the [0, 1] range. The DEV sets from both AfriMTE and our new dataset are used as validation data during training.

LLM-based evaluation We sample the few-shot examples from the training split of the SSA-MTE dataset. For the `por-vmw` language pair, which does not have a training split, demonstrations were instead sampled from the processed and filtered 300 overlapping annotated examples used to assess inter-annotator agreement.

5.1 Model Configurations

We follow the setup of AfriCOMETv1.1 and use the multilingual encoder AfroXLMR-76L, pre-trained on 76 languages widely spoken in Africa. All models are trained using the open-source COMET codebase. Training for the STL and QE models is conducted on a single NVIDIA L40S GPU, while the MTL model is trained on a single NVIDIA A100-SXM4-80GB GPU. We use a batch size of 16 with gradient accumulation over 2 steps. All other hyperparameters follow the default configuration used in AfriCOMETv1.1.

5.2 Baselines

To benchmark the performance of SSA-COMET, we compare it against a wide range of baselines across both MTE and QE settings. These include:

Traditional metrics for MTE *BLEU* and *ChrF++* are lexical overlap metrics based on n-gram precision and character-level F-scores, respectively.

Neural regression-based metrics for MTE For evaluation under the MTE setting, we include *COMET22*, *MetricX-24*, *AfriCOMETv1.0-MTL* (Wang et al., 2024a) (based on AfroX-LMR that supports 20 African languages), *AfriCOMETv1.1-STL* (Wang et al., 2024b) (based on AfroXLMR-76L supporting 76 languages), and *AfriCOMETv1.1-MTL*. The latter is a self-replication model, trained on the same data as *AfriCOMET v1.1-STL* but using a multi-task learning formulation.

Neural regression-based metrics for QE For the QE setting, we evaluate *MetricX-24* and *AfriCOMET v1.1-MTL* in QE mode by disabling the reference input at inference time.

LLM baselines We evaluate four open-weight LLMs such as *Gemma-3 27B-it*, *LLaMA-4 100B*, *LLaMA-4 400B*, and *DeepSeek V3*. Additionally, we conducted an evaluation using some frontier proprietary models such as *GPT-4o (08/24)*, and *Gemini-2.0 Flash*, *Claude-3.7-Sonnet* and *Gemini-2.5 Pro* under both MTE and QE settings as strong prompting-based baselines.

We adopt a **5-shot prompt** setup, guided by the same annotation instructions provided to human annotators. To ensure broad coverage of translation quality levels, we extract the minimum and maximum adequacy scores from the training set and divide the range into five equal intervals. One example is sampled from each interval to construct the 5-shot prompt. The same set of demonstrations is used across all test cases for each language pair to ensure consistency and fairness in evaluation. We experiment with two prompting templates: one

LP	Bleu	ChrF++	COMET22	AfriCOMET v1.1 STL	AfriCOMET v1.0 MTL	AfriCOMET v1.1 MTL	MetricX 24	Claude-3.7	Gemini-pro 2.5	SSA-COMET STL	SSA-COMET MTL
eng-amh	0.352	0.441	0.548	0.588	0.612	0.604	**0.659**	0.566	0.605	0.597	0.629
eng-hau	0.312	0.402	0.405	0.465	0.479	0.476	0.495	0.425	0.471	0.459	**0.502**
eng-kik	0.505	0.599	0.263	0.492	0.556	0.693	0.622	0.696	0.735	0.715	**0.765**
eng-kin	0.392	0.459	0.335	0.507	0.551	0.532	**0.620**	0.536	0.528	0.584	0.602
eng-luo	0.465	0.612	0.361	0.616	0.496	0.693	0.543	0.678	**0.782**	0.689	0.773
eng-twi	0.364	0.502	0.328	0.527	0.537	0.596	0.637	0.652	**0.710**	0.649	0.700
eng-yor	0.382	0.436	0.349	0.442	0.482	0.476	0.455	0.501	0.524	0.489	**0.558**
fra-ewe	0.311	0.426	0.330	0.443	0.494	0.550	0.581	0.614	0.658	0.599	**0.670**
fra-wol	0.476	0.572	0.304	0.493	0.478	0.518	0.560	0.699	**0.750**	0.664	0.732
por-vmw	0.181	0.414	0.198	0.238	0.277	0.237	0.378	0.463	**0.487**	0.280	0.327
Average	0.374	0.486	0.342	0.481	0.496	0.538	0.555	0.583	0.625	0.573	**0.626**

Table 3: **Spearman correlation of MTE metrics with human judgments across LPs**. The best scores are **bolded**.

LP	AfriCOMETv1.1-MTL		MetricX-24		Claude-3.7-Sonnet		Gemini-2.5 Pro		SSA-COMET-QE		SSA-COMET-MTL	
	Spear.	Pear.	Spear.	Pear.	Spear.	Pear.	Spear.	Pear.	Spear.	Pear.	Spear.	Pear.
eng-amh	0.568	0.619	**0.618**	**0.639**	0.541	0.587	0.558	0.593	0.552	0.596	0.591	0.625
eng-hau	0.388	0.405	**0.436**	0.416	0.357	0.382	0.401	0.412	0.390	0.393	0.428	**0.442**
eng-kik	0.655	0.648	0.464	0.452	0.677	0.609	0.703	0.650	0.685	0.657	**0.734**	**0.730**
eng-kin	0.473	0.619	0.592	0.738	0.530	0.694	0.534	0.714	0.511	0.735	**0.561**	**0.788**
eng-luo	0.644	0.638	0.329	0.332	0.672	0.646	**0.757**	0.721	0.653	0.649	0.728	**0.728**
eng-twi	0.561	0.678	0.563	0.686	0.640	0.747	**0.697**	**0.771**	0.614	0.698	0.659	0.753
eng-yor	0.424	0.501	0.405	0.524	0.492	0.595	**0.531**	0.607	0.447	0.547	0.529	**0.617**
fra-ewe	0.483	0.437	0.476	0.430	0.592	0.518	0.623	0.524	0.572	0.533	**0.632**	**0.608**
fra-wol	0.407	0.358	0.291	0.258	0.687	0.623	**0.743**	**0.676**	0.638	0.583	0.689	0.649
por-vmw	0.134	0.168	0.292	0.369	**0.498**	**0.551**	0.481	0.528	0.199	0.281	0.237	0.295
Average	0.474	0.507	0.447	0.484	0.569	0.595	**0.603**	0.620	0.526	0.567	0.579	**0.623**

Table 4: **QE results (Spearman and Pearson correlations) for each LP**. The best scores are **bolded**.

that includes error span detection before adequacy scoring, and one that directly predicts the score without error identification. Full templates for both setups are provided in Figure 6 and Figure 7.

5.3 Main Findings

Superior performance of SSA-COMET in MTE As shown in Table 3, SSA-COMET-MTL achieves the highest average Spearman correlation with human judgments in the MTE setting, outperforming all prior AfriCOMET variants as well as strong prompting-based baselines such as Gemini-2.5 Pro.

Robust QE performance A similar trend is observed in Table 4. Under QE setting, SSA-COMET-MTL ranks first in terms of Pearson correlation and second in Spearman correlation. When excluding the por-vmw language pair, SSA-COMET-MTL achieves the highest average performance across the remaining language pairs.

Gains in Previously Challenging Low-Resource Languages Notably, SSA-COMET shows remarkable improvements on low-resource language pairs where all previous AfriCOMET variants have consistently struggled—particularly on Twi and Wolof. As shown in both Table 3 and Table 4, our model achieves substantial gains in correlation with human judgments for these languages. These results highlight the critical role of in-language, high-quality training data, which allows the model to better capture language-specific characteristics and produce more accurate and reliable quality estimates in low-resource scenarios.

Limitations in Portuguese to Emakhuwa LP Despite strong overall results, SSA-COMET performs relatively poorly on the Emakhuwa (por-vmw) language pair under both MTE and QE settings. Nevertheless, it still outperforms all previous AfriCOMET variants. This underperformance is likely due to the absence of Emakhuwa in the AfroXLMR-76L pretraining corpus, which limits the model's ability to generalize to previously unseen languages. However, LLMs based prompting are not affected by this.

LLM-based prompting is more Robust to the absence of Reference LLMs demonstrate greater robustness to the absence of reference translations. Regression-based metrics achieved worse performance when changing from MTE to QE settings. As shown in Table 3 and Table 4, the drop in Spearman correlation from MTE to QE is relatively small for Claude-3.7 (0.014 on average) and Gemini-2.5 Pro (0.022 on average), in contrast to the obvious declines observed in regression-based models. This indicates that regression models are more depen-

Metric	w/ Error Span	Gemma 3 27B	Llama 4 100B	Llama 4 400B	Deepseek V3 671B	GPT-4o (Aug-2024)	Gemini-2.0 Flash	Claude-3.7 Sonnet	Gemini-2.5 Pro
Spearman	×	0.453	0.446	0.513	0.498	0.506	0.544	0.583	**0.625**
	✓	0.342	0.266	0.484	0.332	0.335	0.508	0.577	**0.590**
Pearson	×	0.498	0.485	0.551	0.530	0.544	0.575	0.609	**0.645**
	✓	0.373	0.269	0.505	0.340	0.361	0.521	0.606	**0.619**

Table 5: **Average correlation performance of LLMs (Spearman and Pearson) across all LPs**, with and without error span annotation prompts. The best scores are **bolded**.

LP	SSA-COMET-STL		SSA-COMET-MTL	
	w/o WMT	w/ WMT	w/o WMT	w/ WMT
eng-amh	0.558	**0.597**	0.587	**0.629**
eng-hau	0.442	**0.459**	0.425	**0.502**
eng-kik	**0.716**	0.715	0.757	**0.765**
eng-kin	**0.592**	0.584	0.554	**0.602**
eng-luo	**0.697**	0.689	0.736	**0.773**
eng-twi	0.645	**0.649**	0.662	**0.700**
eng-yor	**0.494**	0.489	0.530	**0.558**
fra-ewe	**0.658**	0.599	0.638	**0.670**
fra-wol	**0.682**	0.664	0.705	**0.732**
por-vmw	**0.288**	0.280	0.262	**0.327**
Average	**0.577**	0.573	0.586	**0.626**

Table 6: **Spearman correlations for SSA-COMET in STL and MTL setting**–trained with and without WMT data. The best scores are **bolded**.

dent on the presence of reference translations compared to LLMs. Despite the impressive LLM performance, their performance is significantly worse results if we do not provide in-context examples (5-shots) as shown in Appendix H.

Impact of Error Span Prediction on LLMs Table 5 presents a comparison of LLM performance with and without error span prediction. We observe a consistent decline in both Spearman and Pearson correlations when models are prompted to identify error spans prior to generating adequacy scores. For example, Gemini-2.5 Pro's Spearman correlation drops from 0.625 to 0.590, and its Pearson correlation decreases from 0.645 to 0.619. Overall, prompting for error spans before generating the final score does not appear to improve the quality of final predictions. We provide some **qualitative analysis** for Yorùbá showing that the predicted spans are often reliable in Appendix J. Further investigation is still needed to show how useful the predictions are to users of various MT systems.

5.4 Ablation: Impact of WMT Data

Table 6 presents a performance comparison of models trained with and without WMT Non-African data augmentation. As shown, incorporating WMT data yields notable gains in the MTL setting, whereas its impact in the STL setting is comparatively limited. Notably, our annotated SSA-MTE dataset proves highly effective: the model trained solely on SSA-MTE achieves an average Spearman correlation of 0.586 under the MTL setup, already outperforming all AfriCOMET baselines (as shown in Table 3). This highlights the quality and utility of our in-domain annotations, demonstrating that strong performance can be attained even without external training data.

6 Conclusion

In this work, we present SSA-MTE, a high-quality dataset for MT evaluation in Sub-Saharan African languages, covering 13 language pairs and over 63,000 human annotations. Built on this dataset, we introduce SSA-COMET and SSA-COMET-QE for MTE and QE tasks tailored to the low-resource African languages. In our evaluation, SSA-COMET-MTL achieves the highest average correlation with human judgments in MTE, surpassing all prior regression-based metrics and performing competitively with the strong LLM baseline, *Gemini-2.5 Pro*.

To our best knowledge, we are among the first to show that LLM prompting with just five demonstrations can yield strong evaluation performance for under-resourced languages, offering a simple and effective solution. However, it is not efficient. SSA-COMET offers a compelling solution for both MTE and QE senarios, achieving significantly higher efficiency by several orders of magnitude in inference cost (e.g., time and computational resources), while maintaining strong effectiveness when the African language is supported by the pretrained encoder. All data, models, and code are released under open licenses (CC BY 4.0) to facilitate future research and encourage the development of inclusive, regionally adapted, and reliable evaluation tools for African languages.

Limitations

While our work has made significant progress in MT evaluation for African languages, several limitations remain.

The effectiveness of SSA-COMET models remains influenced by the language coverage of the underlying multilingual encoder. For languages such as Emakhuwa, which are not included in the pretraining corpus of AfroXLMR-76L, performance is still limited.

Moreover, our current evaluation primarily focuses on the adequacy dimension of translation quality. Future work could extend this framework to include complementary aspects such as fluency, grammaticality, terminology consistency, and discourse-level coherence, as these factors are especially important in high-stakes or professional translation scenarios.

It is worth noting that, in contrast to the findings of Wang et al. (2024a), this work reveal a relatively small performance gap between reference-based MTE models and reference-free QE models (see Tables 3 and 4). This observation prompts a research question: *as pretrained language models continue to improve in multilingual capabilities, to what extent is the presence of a reference still necessary for reliable translation evaluation?* We leave this investigation for future work.

Ethical Considerations

We employed paid annotators for this project, and paid them appropriate renumeration for their work. We pay each annotator who contributed 3,300 annotations around \$590, while a single translator earned \$700 for the translation of 1,500 sentences. When two translators are available, they earn half of the amount. We do not have other ethical issues with the source of the texts used for translation and annotation, and do not foresee any privacy issues since the source texts are from the general domain—*news domain*.

For the paper writing, ChatGPT is used only for grammar and typo errors check.

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Oluwadara Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Ijeoma Chukwuneke, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson Koffi Kalipe, Jonathan Mukiibi, Salomon Kabongo Kabenamualu, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, and 8 others. 2025. IrokoBench: A new benchmark for African languages in the age of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2732–2757, Albuquerque, New Mexico. Association for Computational Linguistics.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pretrained 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.

Felermino D. M. A. Ali, Henrique Lopes Cardoso, and Rui Sousa-Silva. 2024. Building resources for emakhuwa: Machine translation and news classification benchmarks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14842–14857, Miami, Florida, USA. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

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.

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*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav

Chaudhary, and 1 others. 2021. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1):4839–4886.

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

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

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

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

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, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672.

Chester Palen-Michel, June Kim, and Constantine Lignos. 2022. Multilingual open text release 1: Public domain news in 44 languages. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2080–2089, Marseille, France. European Language Resources Association.

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.

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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

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

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.

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, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Tosin Adewumi, Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Anuoluwapo Aremu, Jessica Ojo, and 39 others. 2024a. AfriMTE and AfriCOMET: Enhancing COMET to embrace under-resourced African languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5997–6023, Mexico City, Mexico. Association for Computational Linguistics.

Jiayi Wang, David Ifeoluwa Adelani, and Pontus Stenetorp. 2024b. Evaluating WMT 2024 metrics shared task submissions on AfriMTE (the African challenge set). In *Proceedings of the Ninth Conference on Machine Translation*, pages 505–516, Miami, Florida, USA. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

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*.

LP	Bleu	ChrF++	COMET22	AfriCOMET v1.1 STL	AfriCOMET v1.0 MTL	AfriCOMET v1.1 MTL	Metric-X 24	Claude-3.7	Gemini-2.5 Pro	SSA-COMET STL	SSA-COMET MTL
eng-amh	0.311	0.446	0.550	0.622	0.645	0.651	**0.671**	0.602	0.636	0.627	0.660
eng-hau	0.322	0.421	0.407	0.474	0.482	0.481	0.506	0.445	0.474	0.467	**0.514**
eng-kik	0.454	0.586	0.259	0.495	0.529	0.688	0.597	0.638	0.685	0.693	**0.764**
eng-kin	0.347	0.500	0.360	0.585	0.701	0.662	0.752	0.698	0.707	0.766	**0.798**
eng-luo	0.408	0.590	0.368	0.604	0.501	0.535	0.535	0.648	0.758	0.680	**0.770**
eng-twi	0.283	0.496	0.444	0.628	0.634	0.698	0.723	0.748	**0.779**	0.728	0.776
eng-yor	0.331	0.455	0.378	0.497	0.583	0.540	0.572	0.591	0.600	0.573	**0.640**
fra-ewe	0.201	0.346	0.307	0.418	0.479	0.514	0.511	0.544	0.569	0.542	**0.644**
fra-wol	0.399	0.568	0.331	0.456	0.475	0.474	0.541	0.649	0.707	0.612	**0.709**
por-vmw	0.160	0.437	0.253	0.305	0.350	0.273	0.478	0.526	**0.535**	0.369	0.410
Average	0.322	0.485	0.366	0.508	0.538	0.567	0.589	0.609	0.645	0.606	**0.668**

Table 7: Pearson correlation of MTE metrics across language pairs. The best scores are **bolded**.

LP	SSA-COMET-STL		SSA-COMET-MTL	
	w/o WMT	w/ WMT	w/o WMT	w/ WMT
eng-amh	0.590	0.627	0.615	**0.660**
eng-hau	0.455	0.467	0.426	**0.514**
eng-kik	0.708	0.693	0.750	**0.764**
eng-kin	0.743	0.766	0.777	**0.798**
eng-luo	0.690	0.680	0.733	**0.770**
eng-twi	0.714	0.728	0.748	**0.776**
eng-yor	0.579	0.573	0.603	**0.640**
fra-ewe	0.621	0.542	0.600	**0.644**
fra-wol	0.637	0.612	0.673	**0.709**
por-vmw	0.358	0.369	0.317	**0.410**
Average	0.609	0.606	0.624	**0.668**

Table 8: Pearson correlations for SSA-COMET-STL and SSA-COMET-MTL trained with and without WMT data. The best scores are **bolded**.

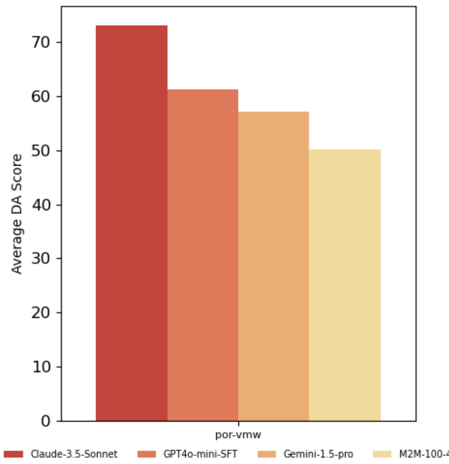

Figure 3: Translation performance of MT systems used for por-vmw.

A Correlations between number of errors and the final scores

Table 9 presents the correlation between Z-normalized DA scores and the frequency of different error types. Among all error categories, mistranslation shows the strongest negative correlation with overall adequacy (Spearman: ˘0.521), followed by addition and omission errors. The aggregated total error count exhibits the highest overall correlation (Spearman: –0.574), confirming that as the number of annotated errors increases, the adequacy score consistently decreases. These findings validate the reliability of error span annotations as strong indicators of perceived translation quality.

Criterion	Z-score	
	Spearman	Kendall
Mistranslation	-0.521	-0.377
Omission	-0.265	-0.210
Addition	-0.276	-0.218
Untranslated	-0.048	-0.038
Total Error	**-0.574**	**-0.406**

Table 9: Correlation of each error criterion with Z-scores.

B Results on AfriMTE

To evaluate the generalization capability of our SSA-COMET models beyond the newly collected SSA-MTE dataset, we conduct experiments on the AfriMTE benchmark (Wang et al., 2024a). As shown in Table 10 and Table 11, SSA-COMET-MTL outperforms all previous AfriCOMET variants, including the strongest one, AFRICOMET-V1.1-MTL. These results demonstrate that SSA-COMET models remain robust and effective under domain shift.

C Data Collection Process for the Portuguese Texts

The Portuguese sentences were sourced from the Multilingual Open Text dataset (Palen-Michel et al., 2022), which features news articles published by Voice of America (VOA[8]). These sentences were translated into Emakhuwa, resulting in a parallel corpus that was released under a CC BY 4.0 license and made publicly available in Ali et al. (2024).

[8] https://www.voanews.com/

LP	AfriCOMET v1.1 STL	AfriCOMET v1.1 MTL	SSA-COMET STL	SSA-COMET MTL
ary-fra	0.526	**0.561**	0.499	0.554
eng-arz	0.510	0.579	0.479	**0.582**
eng-fra	0.492	0.507	0.494	**0.526**
eng-hau	0.561	0.614	0.575	**0.617**
eng-ibo	0.522	**0.582**	0.537	0.564
eng-kik	0.430	0.520	0.415	**0.535**
eng-luo	0.325	**0.515**	0.364	0.506
eng-som	0.502	**0.525**	0.497	0.523
eng-swh	0.704	0.756	0.719	**0.789**
eng-twi	**0.222**	0.209	0.192	0.194
eng-xho	0.203	0.157	**0.233**	0.163
eng-yor	0.338	0.473	0.325	**0.507**
yor-eng	0.508	**0.566**	0.473	0.538
Average	0.449	0.505	0.446	**0.507**

Table 10: Spearman correlation of AfriCOMET and SSA-COMET on AfriMTE. The best scores are **bolded**.

LP	AfriCOMET v1.1 STL	AfriCOMET v1.1 MTL	SSA-COMET STL	SSA-COMET MTL
ary-fra	0.553	**0.641**	0.529	0.627
eng-arz	0.515	**0.603**	0.496	0.593
eng-fra	0.544	0.484	**0.545**	0.500
eng-hau	**0.647**	0.613	0.612	0.637
eng-ibo	0.496	**0.664**	0.520	0.619
eng-kik	0.686	0.545	0.685	**0.696**
eng-luo	0.480	0.526	0.528	**0.624**
eng-som	0.460	0.374	**0.466**	0.393
eng-swh	0.737	0.762	0.745	**0.810**
eng-twi	**0.474**	0.296	0.429	0.457
eng-xho	0.384	0.345	0.376	**0.488**
eng-yor	0.595	0.634	0.592	**0.686**
yor-eng	0.521	**0.571**	0.490	0.531
Average	0.545	0.543	0.539	**0.589**

Table 11: Pearson correlation of AfriCOMET-V1.1 and SSA-COMET on AfriMTE. The best scores are **bolded**.

The dataset has three splits, TRAIN, DEV, and TEST, and covers seven topics: politics, economy, culture, sports, health, society, and world news. We only focus on the annotations for the Test split in this study due to constraints of annotation resources.

D Machine Translations for Emakhuwa

We sampled 1,128 parallel sentences from the Test split of the Portuguese–Emakhuwa dataset. The source sentences were used to generate translations from Portuguese into Emakhuwa using the machine translation systems in Figure 3.

E Emakhuwa Data annotation process

For the more challenging Portuguese-source language pairs, `por-vmw`, we annotated 1,600 samples evenly distributed between 2 evaluators, with 300 overlapping samples split between two evaluators for quality control.

F Further selection of Training Data for Zulu and Igbo

We hypothesize that the low agreement may be due to one evaluator consistently outperforming the other in annotation quality. To address this, we retained only the annotations from the more reliable evaluator for inclusion in the training set. Building on the success of AfriCOMET (Wang et al., 2024a), we employed an AfriCOMET model trained using a multi-task learning framework (Wang et al., 2024a; Wan et al., 2022) on eight language pairs that achieved both Spearman and ICC scores above 0.5 in Table 1. We then used this model to generate predicted scores for `eng-ibo`, `eng-swa`, and `eng-zul`, which served as a silver reference for evaluating annotator reliability. Next, we compared the model-generated scores with those from each evaluator individually and computed both Spearman rank and Pearson correlation coefficients. The results, presented in Table 13, reveal clear gaps in correlation: evaluator 1 for Zulu and Igbo, and evaluator 2 for Swahili, consistently show higher agreement with the silver reference. Therefore, we include their annotations for `eng-ibo` and `eng-zul` in the training set of SSA-MTE.

G Handling for unexpected outputs from LLMs

For a small number of cases, the LLMs fail to generate a valid answer and instead return an un-

Metric	Gemini-2.0 Flash 0-shot	Gemini-2.0 Flash 5-shot	LLaMA4 400B 0-shot	LLaMA4 400B 5-shot	Claude-3.7 0-shot	Claude-3.7 5-shot
Spearman	0.468	0.544	0.325	0.513	0.470	0.583
Pearson	0.506	0.575	0.368	0.551	0.499	0.609

Table 12: Performance differences of LLMs in Zero-shot vs. 5-shot prompting on SSA-MTE.

LP	Evaluator 1		Evaluator 2	
	Spear.	Pear.	Spear.	Pear.
eng-ibo	0.392	0.447	0.277	0.357
eng-zul	0.321	0.363	0.273	0.341

Table 13: Per-annotator Spearman-rank and Pearson correlations with silver references produced by AfriCOMET trained with 8 LPs.

interpretable response. Since our evaluation operates within a normalized range of [0,1], we assign a default score of 0.5—representing a neutral judgment—for these cases. This approach ensures that failing cases do not disproportionately affect overall results, while preserving the integrity of the evaluation. Discarding such cases could introduce selection bias, obscure model weaknesses, and compromise comparability across systems.

H More Details on the Prompting

For all prompting experiments, we used the default decoding settings provided by the API of each LLM. We did not enforce greedy decoding or adjust temperature, top-p, or other sampling parameters. This ensures the results reflect realistic usage scenarios, where users rely on default behavior without fine-tuning generation strategies.

For the 0-shot prompting setup, we removed all demonstration-related content from the prompt, leaving only the annotation guideline and the final instruction for predicting the adequacy score.

Table 12 compares zero-shot and few-shot results, the results shows that without demonstration examples, the performance of the LLMs are unreliable, and far below the performance of SSA-COMET models.

I Comparison: Gemini-2.5 Pro vs. SSA-COMET-MTL

Under the MTE setting, Gemini-2.5 Pro and SSA-COMET-MTL achieve similar overall Spearman correlation. However, when excluding the por-vmw language pair, which is not covered in the pretraining data of the encoder used in SSA-COMET—SSA-COMET-MTL demonstrates a clear advantage, with an average Spearman score that is 0.019 higher. This margin of improvement is comparable to the performance gap between Gemini-2.5 Pro with and without reference input.

Moreover, even when including por-vmw, SSA-COMET-MTL clearly outperforms Gemini-2.5 Pro in terms of Pearson correlation, with a margin of 0.023. This indicates that SSA-COMET-MTL produces adequacy scores that are more accurately aligned with human ratings in absolute terms, not just in relative ranking.

Under the QE setting, excluding the por-vmw language pair from the average, SSA-COMET-MTL achieves a slightly higher Spearman correlation and a notably stronger Pearson correlation, with an advantage of 0.0302.

These results suggest that for languages covered by the encoder, SSA-COMET-MTL is not only more accurate but also significantly more efficient than Gemini-2.5 Pro. On the SSA-MTE test set, SSA-COMET evaluates each LP in under two minutes, whereas prompting LLMs requires substantially more time per sample. This makes SSA-COMET a more scalable and practical solution for low-resource MT evaluation.

J Qualitative evaluation of LLM Error-Span Predictions

Table 18 shows three examples of the predictions of Gemini-2.5 Pro and Llama 4 400 B. We find that the former aligns more with the human judgements than the latter, which aligns with our prompting results in Table 5. Furthermore, we find the error span predictions to be helpful in many cases. We leave a more detailed investigation for future work.

LP	Gemma3-27B-it	Llama4 100B	Llama4 400B	Deepseek V3	GPT4o	Gemini-2.0 Flash	Claude-3.7	Gemini-2.5 Pro
eng-amh	0.503	0.391	0.505	0.496	0.429	0.513	0.566	0.605
eng-hau	0.337	0.323	0.333	0.373	0.411	0.299	0.425	0.471
eng-kik	0.573	0.566	0.666	0.639	0.635	0.687	0.696	0.735
eng-kin	0.467	0.426	0.455	0.461	0.498	0.492	0.536	0.528
eng-luo	0.489	0.524	0.610	0.549	0.639	0.699	0.678	0.782
eng-twi	0.504	0.532	0.578	0.578	0.571	0.613	0.652	0.710
eng-yor	0.399	0.422	0.435	0.434	0.421	0.375	0.501	0.524
fra-ewe	0.430	0.434	0.539	0.461	0.461	0.621	0.614	0.658
fra-wol	0.509	0.519	0.645	0.647	0.663	0.712	0.699	0.750
por-vmw	0.315	0.322	0.363	0.275	0.330	0.431	0.463	0.487
Average	0.453	0.446	0.513	0.498	0.506	0.544	0.583	0.625

Table 14: Spearman correlation of different LLM-based metrics across LPs without generating error spans.

LP	Gemma3-27B-it	Llama4 100B	Llama4 400B	Deepseek V3	GPT4o	Gemini-2.0 Flash	Claude-3.7	Gemini-2.5 Pro
eng-amh	0.556	0.456	0.551	0.551	0.488	0.576	0.602	0.636
eng-hau	0.348	0.350	0.398	0.390	0.421	0.338	0.445	0.474
eng-kik	0.518	0.491	0.608	0.530	0.551	0.607	0.638	0.685
eng-kin	0.648	0.594	0.662	0.649	0.675	0.679	0.698	0.707
eng-luo	0.473	0.498	0.584	0.517	0.595	0.650	0.648	0.758
eng-twi	0.638	0.664	0.685	0.688	0.702	0.733	0.748	0.779
eng-yor	0.553	0.552	0.571	0.576	0.569	0.509	0.591	0.600
fra-ewe	0.382	0.386	0.473	0.462	0.421	0.523	0.544	0.569
fra-wol	0.438	0.432	0.555	0.542	0.591	0.624	0.649	0.707
por-vmw	0.424	0.424	0.423	0.398	0.429	0.512	0.526	0.535
Average	0.498	0.485	0.551	0.530	0.544	0.575	0.609	0.645

Table 15: Pearson correlation of different LLM-based metrics across LPs without generating error spans.

LP	Gemma3-27B-it	Llama4 100B	Llama4 400B	Deepseek V3	GPT-4o	Gemini-2.0 Flash	Claude-3.7	Gemini-2.5 Pro
eng-amh	0.359	0.261	0.486	–	0.335	0.472	0.527	0.598
eng-hau	0.267	0.152	0.310	0.167	0.271	0.289	0.407	0.450
eng-kik	0.422	0.425	0.645	0.539	0.455	0.691	0.702	0.707
eng-kin	0.497	0.281	0.514	0.339	0.465	0.540	0.500	0.529
eng-luo	0.240	0.233	0.538	0.260	0.357	0.615	0.739	0.712
eng-twi	0.384	0.353	0.542	0.415	0.415	–	0.662	0.649
eng-yor	0.373	0.237	0.452	0.396	0.343	0.450	0.470	0.522
fra-ewe	0.334	0.210	0.489	0.307	0.221	0.590	0.581	0.631
fra-wol	0.341	0.347	0.578	0.463	0.422	0.671	0.704	0.724
por-vmw	0.205	0.158	0.284	0.101	0.066	0.256	0.474	0.383
Average	0.342	0.266	0.484	0.332	0.335	0.508	0.577	0.590

Table 16: Spearman correlation of LLM-based metrics across language pairs, using error span prediction. "-" indicates that the model's output failed or collapsed for that LP.

LP	Gemma3-27B-it	Llama4 100B	Llama4 400B	Deepseek V3	GPT-4o	Gemini-2.0 Flash	Claude-3.7	Gemini-2.5 Pro
eng-amh	0.382	0.296	0.503	–	0.333	0.514	0.568	0.611
eng-hau	0.262	0.145	0.353	0.157	0.280	0.322	0.434	0.456
eng-kik	0.400	0.404	0.599	0.472	0.427	0.629	0.651	0.706
eng-kin	0.584	0.283	0.607	0.356	0.529	0.628	0.642	0.692
eng-luo	0.258	0.243	0.531	0.278	0.363	0.591	0.721	0.713
eng-twi	0.459	0.357	0.609	0.448	0.473	–	0.741	0.695
eng-yor	0.493	0.293	0.523	0.488	0.420	0.545	0.594	0.588
fra-ewe	0.303	0.160	0.459	0.271	0.233	0.545	0.537	0.592
fra-wol	0.335	0.346	0.527	0.438	0.407	0.625	0.652	0.702
por-vmw	0.252	0.161	0.336	0.155	0.141	0.292	0.527	0.441
Average	0.373	0.269	0.505	0.340	0.361	0.521	0.606	0.619

Table 17: Pearson correlation of LLM-based metrics across language pairs, using error span prediction. "-" indicates that the model's output failed or collapsed for that LP.

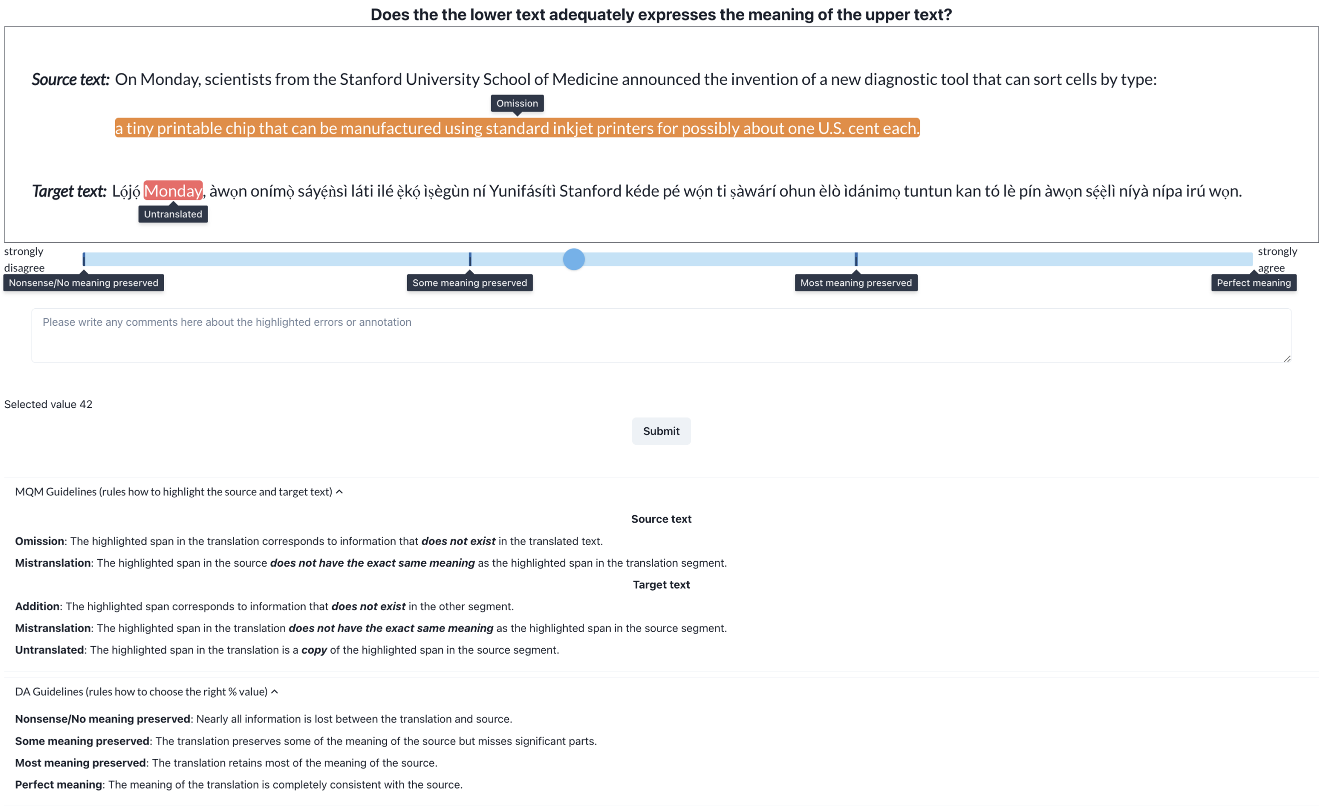

Figure 4: The annotation tool we used for the annotation process.

Meta-Prompt for Prompting LLMs for Translations

Instruction: Translate the following text from <Source Language> to <Target Language>. Return only the translation without any additional text.
Source text: <Source Text> Translation:

Figure 5: The prompt template used for prompting LLMs for translations.

Meta-Prompt for Prompting LLMs with Error Span for MTE

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 corresponding 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.

Instruction: Using the provided source and reference sentences, assess the quality of the machine translation from <Source language> to <Target language> on a continuous scale from 0 to 1, where a higher score indicates better translation quality. Please detect the word-level translation errors before giving the score.
Given examples:

Example 1:
Source: <Example 1 Source Text> Translation: <Example 1 Machine Translation> Reference: <Example 1 Reference Translation>
Output:
The following errors are detected:
<Example 1 Error Spans>
Based on the n error detected, the score of translation is: <Example 1 Score>

…….

Example 5:
Source: <Example 5 Source Text> Translation: <Example 5 Machine Translation> Reference: <Example 5 Reference Translation>
Output:
The following errors are detected:
<Example 5 Error Spans>
Based on the n error detected, the score of translation is: <Example 5 Score>

Based on the examples given, generate the output in exactly the same format, give the error spans and the score, do not give any commentary response.
Source: < Source Text >
Translation: < Machine Translation >
Reference: < Reference Translation>
Output:

Figure 6: The prompt template used for prompting LLMs with error span detection for MTE.

Meta-Prompt for Prompting LLMs without Error Span Detection for MTE

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.

Instruction: Please assess the given machine translation based on the source sentence. Note that you should only output the final score
Given examples:

Example 1:
Source: < Example 1 Source Text > Translation: < Example 1 Machine Translation > Reference: < Example 1 Reference Translation> Score: < Example 1 Score>

…….

Example 5:
Source: < Example 5 Source Text > Translation: < Example 5 Machine Translation > Reference: < Example 5 Reference Translation> Score: < Example 5 Score>

Based on the examples given, generate the output in exactly the same format, give the score and do not give any commentary response.

Source: < Source Text > Translation: < Machine Translation> Reference: < Reference Translation> Score:

Figure 7: The prompt template used for prompting LLMs without error span detection for MTE.

Meta-Prompt for Prompting LLMs without Error Span Detection for QE

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.

Instruction: Please assess the given machine translation based on the source sentence. Note that you should only output the final score
Given examples:

Example 1:
Source: < Example 1 Source Text > Translation: < Example 1 Machine Translation > Score: < Example 1 Score>

…….

Example 5:
Source: < Example 5 Source Text > Translation: < Example 5 Machine Translation > Score: < Example 5 Score>

Based on the examples given, generate the output in exactly the same format, give the score and do not give any commentary response.

Source: < Source Text > Translation: < Machine Translation> Score:

Figure 8: The prompt template used for prompting LLMs without error span detection for QE.

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 **corresponding 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 9: The annotation guideline we used for the annotation process.

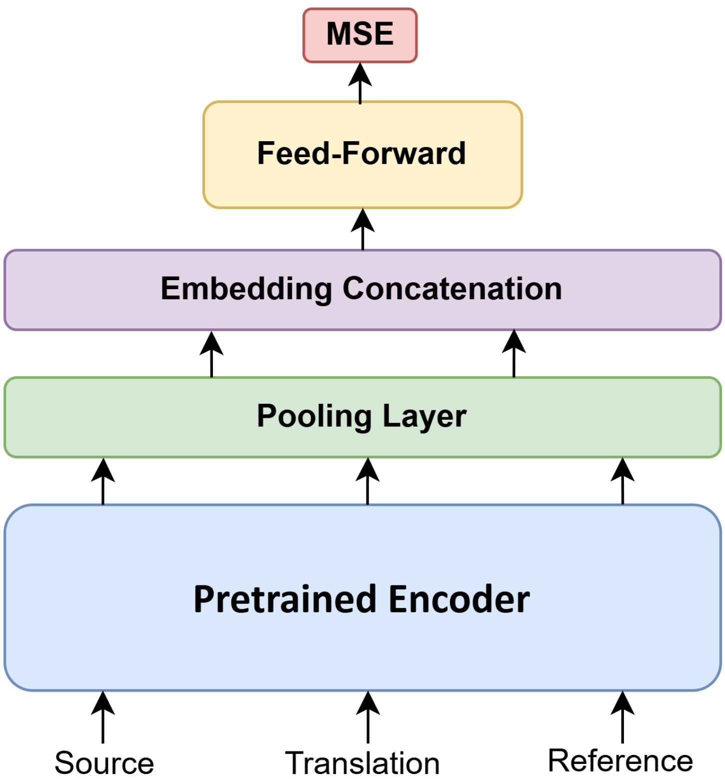

Figure 10: The workflow of the COMET architecture

Sentence / LLM output	Human rating
Source: Angola, as everyone knows, is a beautiful country. **Translation:** Gẹgẹbi gbogbo eniyan ti mọ, Angola jẹ orilẹ-ede ti o dara julọ. **Translation (in English)** As everyone knows, Angola is the most beautiful country. **Reference:** Angola, bí gbogbo ènìyàn ṣe mọ̀, jẹ́ orílẹ̀-èdè tí ó rẹwà.	0.77
Gemini 2.5 Pro Output **Output:** The following error is detected: **Mistranslation:** The span "beautiful" in the source text was wrongly translated into "dara julọ" in the translation. **Source:** Angola, as everyone knows, is a <s>beautiful</s> country. **Target:** Gẹgẹbi gbogbo eniyan ti mọ, Angola jẹ orilẹ-ede ti o <s>dara julọ</s>. Based on the 1 error detected, **the score of translation is: 0.67**	
LLama 4 400B Output The following error is detected: **Mistranslation:** The span "beautiful country" in the source text was wrongly translated into "orilẹ-ede ti o dara julọ" in the translation. **Source:** Angola, as everyone knows, is a <s>beautiful country. </s> **Target:** Gẹgẹbi gbogbo eniyan ti mọ, Angola jẹ <s> orilẹ-ede ti o dara julọ</s>. Based on the 1 error detected, **the score of translation is: 0.58**	
Source: Nobody knew Alffa outside a small part of Wales **Translation:** Kò sí ẹni tó mọ Alffa ní àgbègbè kékeré kan ní Wales **Translation (in English)** Nobody knew Alffa in this small area of Wales **Reference:** Kò sí ẹni tó mọ Alffa yàtọ̀ sí àwọn ibi kéréje ní Wales	0.76
Gemini 2.5 Pro Output **Output:** The following error is detected: **Mistranslation:** The span ""outside"" in the source text was wrongly translated into ""ní"" in the translation. **Source:** Nobody knew Alffa <s>outside</s> a small part of Wales **Target:** Kò sí ẹni tó mọ Alffa <s>ní</s> àgbègbè kékeré kan ní Wales Based on the 1 error detected, **the score of translation is: 0.67**	
LLama 4 400B Output The following error is detected: **Mistranslation:** The span ""outside a small part of Wales"" in the source text was wrongly translated into ""ní àgbègbè kékeré kan ní Wales"" in the translation. **Source:** Nobody knew Alffa <s>outside a small part of Wales</s> **Target:** Kò sí ẹni tó mọ Alffa <s>ní àgbègbè kékeré kan ní Wales </s> Based on the 1 error detected, **the score of translation is: 0.38**	
Source: However, professional freelance photojournalists are still quite rare. **Translation:** Sibẹsibẹ, awọn onise iroyin ti o ni imọ-ẹrọ ti o ni imọ-ẹrọ ti o ni imọ-imọ-imọ-imọ-imọ-imọ-iimọ-imọ-imọ-imọ-imọ-imọ-... **Translation (in English)** However, the news staffs that have technical knowledge that have technical knowledge knowledge knowledge knowledge knowledge knowledge knowledge ... **Reference:** Ṣùgbọ́n, àwọn akọ́ṣẹmọṣẹ́ afàwòránròyìn adádúró kò sì wọ́pọ̀ náà.	0.11
Gemini 2.5 Pro Output **Output:** The following error is detected: **Mistranslation:** the span "professional freelance photojournalists are still quite rare" in the source text was wrongly translated into "wọn onise iroyin ti o ni imọ-ẹrọ ti o ni imọ-ẹrọ ti o ni imọ-imọ-imọ-imọ-imọ-imọ-iimọ-imọ-imọ-imọ-imọ-..." in the translation. **Source:** However, <s>professional freelance photojournalists are still quite rare</s> **Target:** Sibẹsibẹ <s>awọn onise iroyin ti o ni imọ-ẹrọ ti o ni imọ-ẹrọ ti o ni imọ-imọ-imọ-imọ-imọ-imọ-iimọ-imọ-imọ-imọ-imọ-imọ-...</s> Based on the 1 error detected, **the score of translation is: 0.16**	
LLama 4 400B Output The following error is detected: **Mistranslation:** The span "professional freelance photojournalists" in the source text was wrongly translated into "wọn onise iroyin ti o ni imọ-ẹrọ ti o ni imọ-ẹrọ ti o ni imọ-imọ-imọ-imọ-imọ-imọ-iimọ-imọ-imọ-imọ-imọ-imọ-..." in the translation. **Source:** However, <s>professional freelance photojournalists</s> are still quite rare. **Target:** Sibẹsibẹ <s>awọn onise iroyin ti o ni imọ-ẹrọ ti o ni imọ-ẹrọ ti o ni imọ-imọ-imọ-imọ-imọ-imọ-iimọ-imọ-imọ-imọ-imọ-imọ-...</s> Based on the 1 error detected, **the score of translation is: 0.0**	

Table 18: Examples of Error-Span Predictions by different LLMs, ratings are from 0 to 1. The mistakes of the translation model is in red.