Trust but Check: LLM-Assisted Review of Human Translations in African Languages

Tadesse Destaw Belay[1], **Henok Biadglign Ademtew**[2], **Idris Abdulmumin**[3], **Sukairaj Hafiz Imam**[4], **Abubakar Juma Chilala**[5], **Godfred Agyapong**[6], **Chinedu Emmanuel Mbonu**[7,8], **Basil Friday Ovu**[9], **Catherine Nana Nyaah Essuman**[10], **Alfred Malengo Kondoro**[11], **Sonia Adhiambo**[12], **Daud Abolade**[12,13], **Ponts'o Mpholle**[12], **Nicholaus Ladislaus**[5], **Saminu Mohammad Aliyu**[4], **Gali Ahmad Samuel**[13], **Fabrice Hakuzimana**[14], **Mike Nzirainengwe**[12], **Temitayo Olatoye**[15], **Sileshi Bogale Haile**[16], **Tewodros Achamaleh Bizuneh**[1], **Tolulope Olalekan Abiola**[1], **Kedir Yassin Hussen**[17], **Ibrahim Said Ahmad**[18], **Verrah Akinyi Otiende**[12,19,20], **Seid Muhie Yimam**[21], **Shamsuddeen Hassan Muhammad**[4,22]

[1]Instituto Politécnico Nacional, [2]EthioNLP, [3]University of Pretoria, [4]Bayero University Kano, [5]Carnegie Mellon University, [6]University of Florida, [7]Nnamdi Azikiwe University, [8]Nazarbayev University, [9]Alvan Ikoku Federal University of Education, [10]Umbaji SARL, [11]Hanyang University, [12]Masakhane, [13]University of Lagos, [14]Digital Umuganda, [15]University of Eastern Finland, [16]Assosa University, [17]Gondar University, [18]Northeastern University, [19]University of Michigan, [20]USIU-Africa, [21]University of Hamburg, [22]Imperial College London

Contact: tadesseit@gmail.com

Abstract

Large-scale translation projects for low-resource languages mostly rely on human translators to ensure cultural and linguistic fidelity. However, even professionally produced translations often contain subtle translation errors that are difficult to detect. Manual quality control at scale becomes prohibitively expensive, creating a major bottleneck in the development of high-quality Natural Language Processing (NLP) resources. Recent advances in multilingual large language models (LLMs) offer promising support for annotation workflows with human-in-the-loop settings. In this work, we investigate the use of LLMs to assist in auditing translation quality, enabling more efficient quality control pipelines for low-resource African languages. We audit translations in 11 African languages using the MAFAND-MT dataset, combining LLM-as-a-judge, native-speaker human review, and automated metrics. Our quality-audited version of MAFAND-MT test set yields performance gains across all languages, with BLEU scores ranging from 0.4 to 9.27 and chrF scores ranging from 0.3 to 8.69. Our findings further indicate that state-of-the-art LLMs, such as GPT-5.1, can assist in auditing translation quality and suggesting candidate corrections for low-resource languages. However, they remain far from being a stand-alone solution for the automatic correction of human translations in African languages.

1 Introduction

Machine Translation (MT) is a fundamental and prominent task in natural language processing (NLP), essential for global communication and information access (Anastasopoulos et al., 2020). For many low-resource languages, particularly those in Africa, a common method for developing benchmark datasets is through human translation of existing resources from higher-resource languages such as English and French (Adelani et al., 2025a). Therefore, the quality of these translated datasets is very crucial, as it directly impacts the evaluation and development of MT systems, ultimately determining their reliability for real-world use. High-quality human translations should satisfy at least three key criteria: fluency in the target language, adequacy in preserving the semantic content of the source text, and the target language's cultural context (Freitag, Markus and Foster, George and Grangier, David and Ratnakar, Viresh and Tan, Qijun and Macherey, Wolfgang, 2021).

However, human translation, while indispensable for cultural and nuanced understanding, is not immune to error (Han et al., 2021; Lin et al., 2022). Translators may introduce typos, grammatical mistakes, fluency issues, and bilingual (code mixing) errors (Lin et al., 2022). These errors can stem from various factors, including the use of imperfect auxiliary translation tools, errors by native translators, the translator's proficiency in the target language,

and the ambiguity of the source content to be translated (Han et al., 2021; Lin et al., 2022). Table 1 shows examples from the MAFAND-MT test set where human-translated text in Amharic, Hausa, Igbo, Swahili, and Twi languages contains such errors, creating a "garbage in, garbage out" risk for MT models and evaluations (Adelani et al., 2022). Furthermore, errors that propagate into benchmark datasets systematically bias evaluation and hinder the development of robust MT models (Koehn and Knowles, 2017).

Despite their importance, ensuring the quality of human translations at scale remains a major challenge. Exhaustive manual review by professional translators is financially unsustainable (Sambasivan et al., 2021). Consequently, many projects face a difficult trade-off between scale, cost, and quality, potentially allowing errors to propagate into valuable resources.

Recent advances in multilingual Large Language Models (LLMs) offer a promising path toward multidisciplinary problem-solving capabilities (Treviso et al., 2024; Feng et al., 2025a). In this work, we investigate whether LLMs can act as assist *first-pass filters*, automatically identifying translation errors with a higher likelihood of containing errors and thus proceeding for expert review. Specifically, we explore the following three research questions: **RQ1** Can large language models (LLMs) assist in detecting and correcting human translation errors in low-resource African languages (how are they good enough to judge the translation quality of low-resource languages)? **RQ2** What types of translation errors are commonly found in machine translation (MT) resources for African languages? and **RQ3** How does translation quality review improve the performance of machine translation systems in low-resource languages?

Our contributions are threefold: (1) We introduce a pipeline for LLM-assisted quality assurance of translated resources; (2) We provide a detailed error analysis of a subset of human-translated MAFAND-MT datasets (11 languages), (3) We explore the different kinds errors exist in African MT resources; and (4) We offer insights on using LLMs as a cost-effective alternative in the reviewing of a high-quality human translation dataset for low-resource languages. Our findings demonstrate that an LLM-human workflow can help develop reliable and accurate MT datasets and systems.

2 Related Work

2.1 Auditing Training Corpora Quality

Recent efforts to improve NLP for African languages have increasingly emphasized both the scale and quality of training corpora. Early work focused on constructing large-scale web-crawled multilingual pretraining datasets (Xue et al., 2021; Vegi et al., 2022b; Tonja et al., 2024), demonstrating the feasibility of incorporating a broader set of African languages into foundation models. However, the reliance on web-crawled data introduced substantial noise, including mistranslations, misalignments, and non-parallel content, which introduces greater degradation in data quality for low-resource languages.

To mitigate these quality issues, subsequent studies have focused on improving the quality of translation datasets through filtering, cleaning strategies, and manual or semi-automatic audits (Zhang et al., 2020; Kreutzer et al., 2022). This evolution reflects a growing recognition that data quality, rather than volume, is a critical bottleneck for machine translation performance in low-resource settings.

Despite these advances, existing auditing approaches remain largely human-intensive, limiting their scalability across languages and domains. In this work, we position LLMs as a complementary tool for MT data auditing, examining their ability to judge if the given translation is whether correct or not, identify translation errors and suggests correct translations.

2.2 Evaluations of Test Dataset Quality

Language models are commonly evaluated on downstream tasks, with Machine Translation (MT) serving as a central benchmark for assessing cross-lingual capabilities. For African languages, MAFAND-MT (Adelani et al., 2022) is one widely used evaluation dataset that has supported numerous studies on MT training and evaluations (Ojo et al., 2025; Abdulmumin et al., 2022; Vegi et al., 2022a; Nzeyimana, 2024; Tang et al., 2024; Ji et al., 2025; Singh et al., 2025). The validity of conclusions drawn from such benchmarks critically depends on the quality of their underlying translations.

The work by Abdulmumin et al. (2024) demonstrated that even human-translated evaluation datasets are susceptible to translation errors and identified and corrected issues for some African languages (Hausa, Sepedi, Xitsonga, isiZulu) in

MAFAND translated dataset	Corrected translation
"eng": "Date: Thursday, July 31, 2014", "amh": 'የዘጎ ዘጠኝ ጦማሪዎች ፍትህ ይገበቸዋል"	"eng": "Date: Thursday, July 31, 2014", "amh": ቀን፡ሐሙስ፣ ሐምሌ 31፣ 2014
"eng": "Ian Simbota represented the Association of Persons with Albinism at the court.", "kin": "Igihe umucamanza yasomaga, icyampangayikishije ni uko igice cya Padiri Muhosha [n'abandi] bahamwe n'icyaha ..."	"eng": "Ian Simbota represented the Association of Persons with Albinism at the court.", "kin": "Ian Simbota yari ahagarariye ishyirahamwe ry'abantu bafite ubumuga bw'uruhu mu rukiko."
"eng": "Every soul shall have a taste of death.", "hau": "Ko shakka babu akwai wani lokaci da rayuwa za ta zo karshe Ƙowane rai mai dandanar mutuwa neƆSuratu Al Imrana 3:185)."	"eng": "Every soul shall have a taste of death.", "hau": "Ubangiji, muna d⌢aukin jiran wannan damar"
"eng": "policemen has claimed ownership of Dino melaye", "ibo": "Ndị uweojii egbochiela ụlọ Dino Melaye"	"eng": "policemen has claimed ownership of Dino melaye", "ibo": "Ndị uweojii ekwuola na ha nwe Dino Melaye"
"eng": "Ian Simbota represented the Association of Persons with Albinism at the court.", "swa": "Igihe umucamanza yasomaga, icyampangayikishije ni uko igice cya Padiri Muhosha [n'abandi] bahamwe n'icyaha ...."	"eng": "Ian Simbota represented the Association of Persons with Albinism at the court.", "swa": "Ian Simbota yari ahagarariye ishyirahamwe ry'abantu bafite ubumuga bw'uruhu mu rukiko."
"eng": "A local councilor, Jabu Zondo, visited the area yesterday to reprimand the incident.", "twi": "Kurow no mu kaunsila, ɛabu Zondo kɔɔ beaeɛ hɔ nnera kɔhwɛɛ deɛ ɛsii no"	"eng": "A local councilor, Jabu Zondo, visited the area yesterday to reprimand the incident.", "twi": "Kurow no mu gyinatufoɔ panyin, Jabu Zondo, kɔɔ beaeɛ no nnera kɔ kaa wɔn anim."

Table 1: **Examples of translation errors from the MAFAND-MT test dataset.** The Table shows example cases in Amharic (amh) Hausa (hau), Igbo (ibo), Swahili (swa), and Twi (twi) translations (Adelani et al., 2022). The red marked text is the incorrect translation of the given English-sourced (eng) text, and the blue is the correct translation using native speakers of the target language.

the FLORES dataset. These findings, together with prior analyses (Freitag, Markus and Foster, George and Grangier, David and Ratnakar, Viresh and Tan, Qijun and Macherey, Wolfgang, 2021), highlight the need for systematic auditing and validation of MT evaluation datasets to ensure reliable benchmarking.

However, existing approaches primarily rely on additional rounds of human translation and expert review, which are costly and difficult to scale across languages and datasets. In contrast, the use of LLMs for auditing evaluation data quality remains underexplored. In this work, we investigate the extent to which LLMs can assist in auditing MT evaluation datasets by identifying translation errors and inconsistencies, and we analyze their agreement with human judgments to better understand when LLM-based auditing can reduce cost and when human oversight remains essential.

2.3 LLM-as-a-Judge Translation Review

LLMs have demonstrated strong performance as an evaluator in various tasks, including automated data quality control (Gu et al., 2025), dataset annotation assistance (Tan et al., 2024; Belay et al., 2025), identifying error types for machine translation dataset (Feng et al., 2025b; Kim, 2025), and research paper summarization (Liu et al., 2024). Within machine translation research, LLM-assisted translation error detection and correction have been explored primarily in English as an automatic post-editing (APE) (Berger et al., 2024; Freitag, Markus and Foster, George and Grangier, David and Ratnakar, Viresh and Tan, Qijun and Macherey, Wolfgang, 2021; Lu et al., 2024), translation quality evaluation (Qian et al., 2024), and automatic correction of human translations (Lin et al., 2022).

Despite these advances, the use of LLMs as translation reviewers, capable of identifying, categorizing, and correcting translation errors, remains unexplored, mainly for low-resource and African languages. Recent evaluations of MAFAND-MT (Adelani et al., 2022) reveal the presence of translation errors and varying degrees of semantic misalignment, as evidenced by low automatic quality scores reported in prior work using metrics such as COMET (Falcão et al., 2024). These underscore the need for scalable, systematic translation-review methods that extend beyond manual inspection. In this work, we investigate the role of LLMs as translation reviewers for African-language datasets. Specifically, we assess their

ability to judge whether the translation is correct, identify common types of translation errors, propose correction candidates, and support a human-in-the-loop auditing pipeline. By analysing agreement between LLM-based judgments and human verification, we aim to clarify both the potential and the limitations of LLMs as assistants for translation data quality auditing.

3 Translated African Languages Dataset

A growing number of human-translated datasets are available for African languages. These datasets are mostly translated from English and French source texts and span diverse domains. For machine translation NLP tasks, prominent datasets that include African languages are FLORES (Guzmán et al., 2019), NLLB (Team et al., 2022), HornMT[1], and MAFAND-MT (Adelani et al., 2022). Recently, domain-specific datasets have also been created, such as AFRIDOC-MT (Alabi et al., 2025) and AfriMed-QA (Nimo et al., 2025) for health-related data, AfriGSM (Adelani et al., 2025b) for math word problems, and AfriMMLU and AfriXNLI (Adelani et al., 2025b) for general knowledge and reasoning.

The MAFAND-MT Translation Dataset The Masakhane Anglo and Franco Africa News Dataset for Machine Translation (MAFAND-MT) is one of the frequently used MT evaluation datasets for African languages (Adelani et al., 2022). This dataset covers 20 African languages: 15 are translated from English into target languages, and the remaining 5 are translated from French sources. MAFAND-MT data is professionally translated by native speakers of the target languages with a compensation (Adelani et al., 2022). However, we observed a range of common MT error types presented in Table 1.

4 Translation Review Pipeline

We used a two-stage review pipeline to assess and review translation quality. In the first stage, LLM automatically evaluates each translation pair and flags potential errors. In the second stage, native-speaker verify the flagged cases and provide corrections where necessary, shown in Figure 1.

[1]A multi-way parallel news corpus for languages in the Horn of Africa; https://github.com/asmelashteka/HornMT

Category	Error description
Typos	Misspellings or character-level mistakes in the translation
Grammar	Grammatical errors (e.g., agreement, tense, syntax)
Fluency	Unnatural or awkward phrasing; non-native flow
Bilingual	Interference or overly literal translation from English
Incomplete	Translation omits some(all) part(s) of the source meaning
Addition	Adds information not present in the source
Omission	Removes information present in the source

Table 2: **Error categories reported during manual analysis of translation quality.** The table summarizes recurrent error types observed with their description.

LLM-assisted Translation Review We used GPT-5.1[2] as a judge to review the translation. We probed GPT-5.1 for each parallel text pair to assess translation quality and classify the translation as `correct` or `incorrect`. We further instructed this LLM to suggest the types of translation error(s) presented in Table 2 and the correct translation versions if the reply was `incorrect` at first.

Human Translation Correction We assign a minimum of two native-speaker volunteers per language for a total of 11 languages. Human translation reviews translation errors flagged by LLMs as `incorrect` and verifies the corrected translations proposed by the LLMs. To facilitate native speaker review of LLM suggested corrections, we design an interactive annotation interface that displays the source English text, the original MAFAND-MT human translation, and the LLM proposed alternative. If either translation is flagged as incorrect, the tool highlights common error categories. Annotators are also provided with an option to supply a new translation if both existing options are incorrect. Details of the translation review guidelines and the annotation interface are provided in Appendix A.

5 Human vs LLM Audit Agreement

We analyze the agreement between humans and GPT-5.1 in translation quality review and assess whether the LLM's suggested corrections are useful.

Can LLMs assist in reviewing translation quality for low-resource languages? Based on AfroBench, a benchmark for evaluating LLMs on African languages, proprietary LLMs such as the GPT family consistently outperform widely used open-source models for the machine translation task (Ojo et al., 2025). Motivated by this observation, we used the latest GPT-5.1 as a transla-

[2]https://openai.com/index/gpt-5-1/, Dec 2025

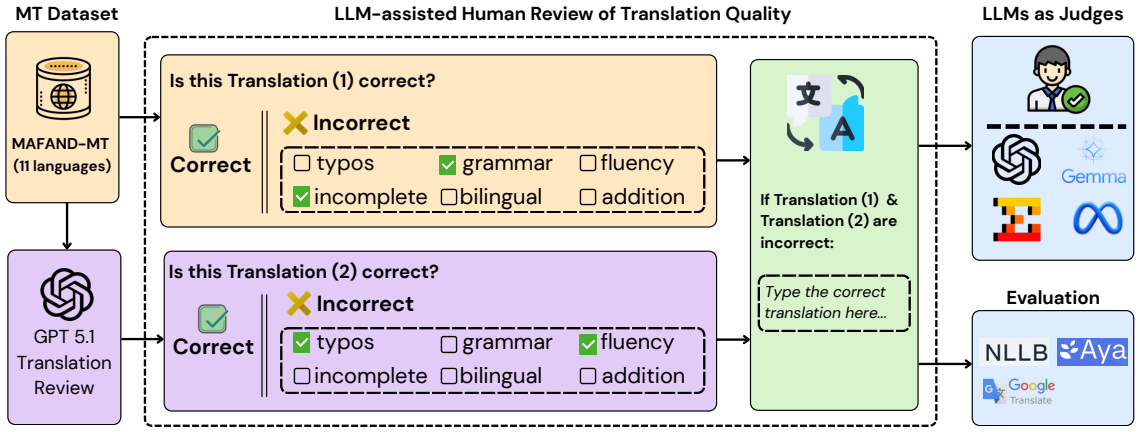

Figure 1: **LLM-assisted pipeline for MT dataset quality assessment and correction.** The figure illustrates the workflow for judging whether the translation is correct or not, identifying translation errors, and generating correction suggestions.

Language	Translation direction	# Test set	LLM predict "Correct"	LLM predict "Incorrect"	Human vs LLM Trans. Agree. %	Human vs LLM Errors. Agree.
Amharic	eng-amh	1,037	271	766	0.85	0.23
Hausa	eng-hau	1,500	680	820	0.23	0.25
Igbo	eng-ibo	1,500	884	964	0.78	0.18
Kinyarwanda	eng-kin	1,006	390	616	0.55	0.00
Luo (Dholuo)	eng-luo	1,500	398	1,102	0.96	0.09
Nigerian Pidgin	eng-pcm	1,564	1009	555	0.43	0.20
Shona	eng-sna	1,005	368	637	0.24	0.15
Swahili (Kiswahili)	eng-swa	1,835	748	787	0.30	0.05
Tswana (Setswana)	eng-tsn	1,500	844	656	0.75	0.18
Twi (Akan-Twi)	eng-twi	1,500	265	1285	0.91	0.11
Yoruba	eng-yor	1,558	585	973	0.45	0.16

Table 3: **The MAFAND-MT test set dataset details with human and LLM agreement analysis**. **LLM predict Correct** and **LLM predict Incorrect** columns are the number of translation responses from LLMs. **Human vs LLM Translation Agreement** is the percentage agreement between humans and LLM to say the original translation is correct or incorrect. **Human vs LLM Translation Agreement** is Cohen's Kappa translation error label agreement between Human and LLM. We target only the test set data, and each language has its own Source (English) and target translations.

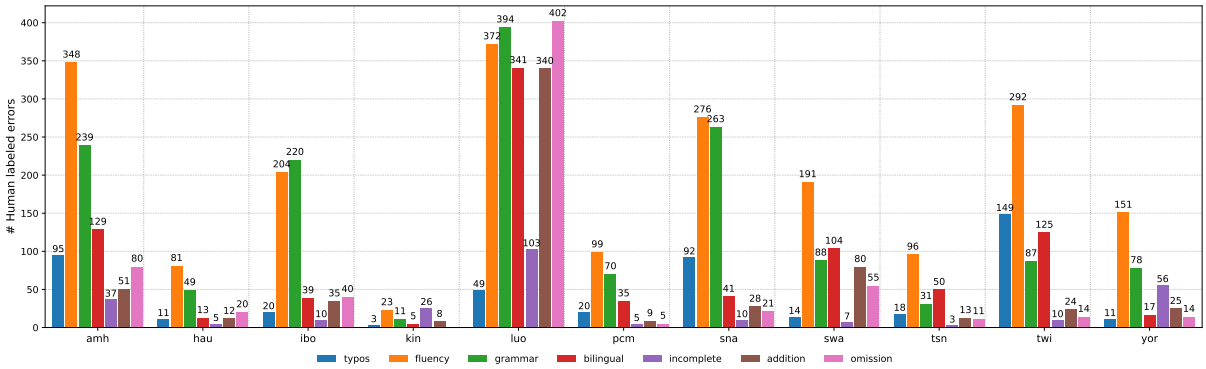

Figure 2: The statistics of Human-labeled translation errors for the MAFAND-MT test dataset. The bar graph illustrates the number of types of translation errors where Amharic (amh), Luo-Dholuo (luo), Igbo (ibo), Shona (sna), Swahili (swa), and Twi (twi) languages have more error types statistically.

tion quality reviewer. Native speakers of each target language are then presented with two options: the original MAFAND-MT translation and the revised translation produced by the GPT-5.1

(reviewer model).

The level of agreement between GPT-5.1 and humans in assessing translation quality is reported in Table 3. The results suggest that GPT-5.1 can support translation quality verification within a human-in-the-loop framework. Notably, based on GPT-generated revisions, a substantial number of translations were judged by native speakers to be of higher quality than the original human-produced translations, for example, amh (118), hau (628), yor (124), swa (548), and pcm (320).

Regarding agreement on translation error labels between the LLM (GPT-5.1) and human native speakers (shown in Table 3, *Human vs. LLM Errors Agreement* column), the Cohen's Kappa scores are consistently low, ranging from 0 (minimum) to 0.25 (maximum). This low agreement can be attributed to several factors: (1) translation errors are annotated in a multi-label setting, (2) GPT-5.1 tends to over-predict multiple error types for a single translation pair, and (3) there is substantial disagreement in cases where native speakers labeled most LLM translations as incorrect.

State-of-the-art LLMs, such as GPT-5.1, can assist with machine translation quality auditing and provide translation suggestions for low-resource languages. A considerable number of translations that were generated by LLM were judged by native speakers to be of acceptable quality. Moreover, a majority of annotators (64.7%) reported that qualifying each translation took 1–3 minutes, and 29.4% reported 30 seconds to 1 minute, indicating that LLM pre-auditing can provide a substantial time-saving benefit for human annotators. Regarding the helpfulness of adding LLM-generated translation as an option during quality audits of machine translation data, native speaker responds 42.2% yes, it was helpful, 42.1% partially useful, and 17.6% not helpful. However, the selected LLM reviewer (GPT-5.1) remains far from a stand-alone solution for the automatic correction of human translations in African languages. High-quality, corrected MAFAND-MT test set data will therefore be valuable for further machine translation experiments and evaluations.

Can LLMs serve as judges of translation quality for low-resource languages? We select the following popular open-source LLMs for the LLM-as-a-judge evaluation: Gemma-3-27B (Team et al., 2025), LLaMA-3.3-70B (Grattafiori et al., 2024), GPT-oss-120B (OpenAI et al., 2025), and Mistral-123B (OpenAI et al., 2024). In addition, we include the closed-source model GPT-5-mini (Hurst et al., 2024). We evaluate these models on 200 randomly sampled, human-labeled instances (100 correct and 100 incorrect translations), where incorrect translations are further annotated with fine-grained error labels.

As shown in Figure 3, the agreement between human judgments and LLM-based judges is substantially low. In particular, for translations labeled as `correct` by human annotators, most LLM judges incorrectly classify them as incorrect. Models such as Mistral-123B, GPT-oss-120B, and GPT-5-mini fail to identify correct translations reliably. A similar trend is observed for `incorrect` translations: the LLM judges tend to label nearly all translation pairs as incorrect and frequently overpredict multiple error categories for a single translation. The agreement on error labeling between humans and LLM-as-judge models is reported in Appendix 6.

How is the COMET score a reliable quality estimation for low-resource languages? We apply SSA-COMET-QE (Sub-Saharan African Cross-lingual Optimized Metric for Evaluation of Translation) - an improved version of AfriCOMET (Wang et al., 2024), a robust and automatic metric for machine translation quality estimation (Rei et al., 2020). It receives a pair (source sentence, translation in target language), and returns a score ranging from 0 (semantically unrelated) to 1 (high quality) that reflects the quality of the translation (Li et al., 2025). The quality estimation score (SSA-COMET-QE) for the MAFAND-MT dataset before and after quality check is presented in Figure 4. Based on the SSA-COMET-QE score, each language has translations ranging from 70 to 400 parallel texts that have a quality estimation score of less than 0.6. As illustrated in Figure 4, an empty source and output or a single character or word translation can still receive a low score. Across all languages, the maximum score observed is 0.84; even perfect translations do not approach a score of 1.0, and most translation outputs fall within a narrow range between 0.65 and 0.70. Notably, some outputs in an incorrect language receive higher scores than null or random outputs produced in the correct language. These findings indicate a divergence between SSA-COMET-QE scores and human judgments in low-resource language settings, warranting further investigation into the reliability and behaviour of such evaluation metrics for low-resource

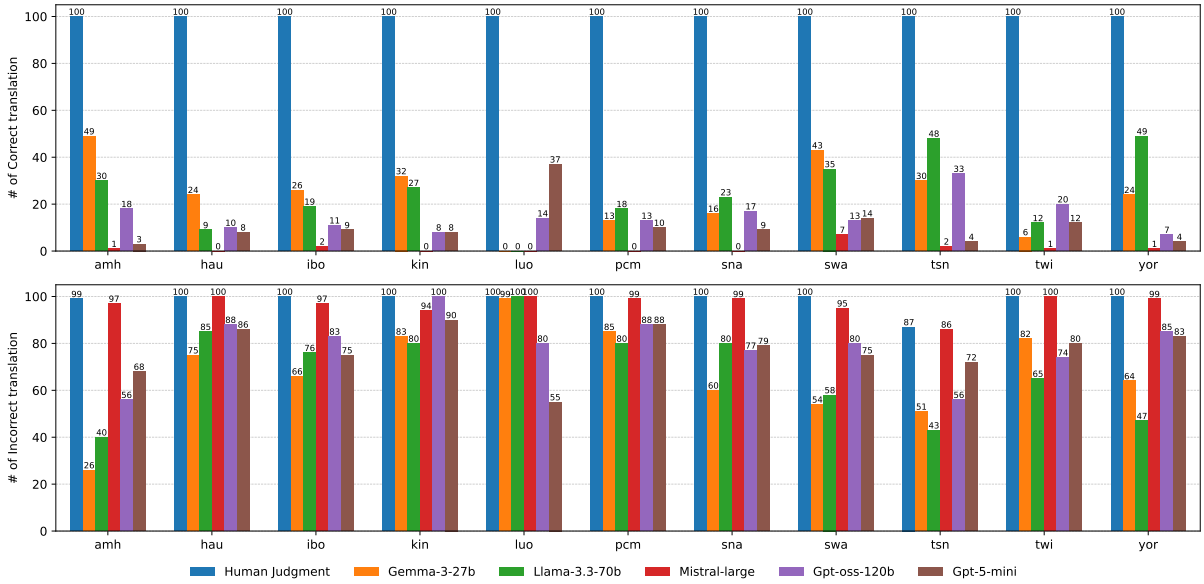

Figure 3: Human judgments versus LLM-as-a-judge on 200 randomly selected samples: 100 translations labeled as `correct` and 100 as `incorrect` by human evaluators.

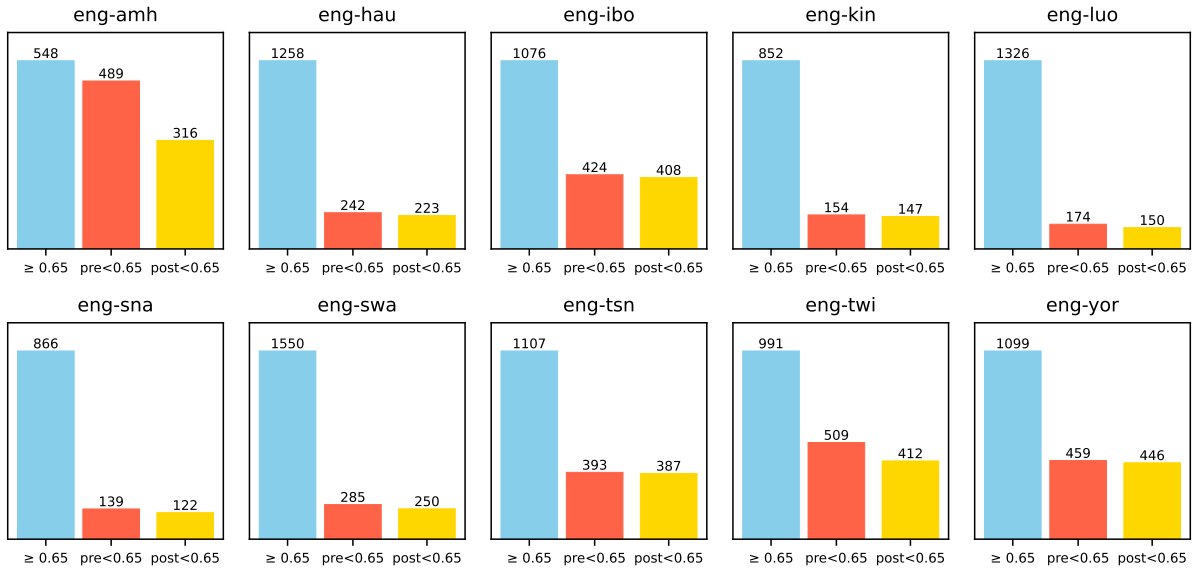

Figure 4: **SSA-COMET-QE translation scores across language pairs before (pre) and after (post) applying quality audit.** Scores range from 0 to 1, with higher values indicating better translation quality. The `en-pcm` direction has lower COMET scores; only 21% of the data have>0.5 SSA-COMET-QE because the language is not included in the specifically fine-tuned model.

languages. While we improved the statistics of low-scoring translations, the improvement remains modest due to the quality of the SSA-COMET-QE metrics, as shown in Table 4.

6 Benchmarking Improved Dataset

Automated scores provide a cost-effective and rapid approximation of quality, which is essential for machine translation system performance and for quick feedback on evolving models (Kocmi et al., 2024). While human judgment remains the gold standard, we evaluate our approach in three ways: human judgments, LLM-as-a-judge assessments, and automatic evaluation metrics such as BLEU and chrF for MT output. To make benchmark results, we select popular open-source machine translation models such as multilingual Aya-101 (Üstün et al., 2024), NLLB 600M and NLLB 3B (Team

Language	Source text	Translation	SSA-COMET score
amh	*NULL*	*NULL*	0.38
amh	1	1	0.44
amh	" "	" "	0.48
ibo	Naira	Naira	0.54
ibo	Ihiala	Ihiala	0.44
swa	ICANNWiki	ICANNWiki	0.49
swa	(CC BY 2.0)	(CC BY 2.0)	0.55
pcm	"GOD DEY,"	"GOD DEY,"	0.33

Table 4: Translation pair examples with SSA-COMET-QE score. Short translation examples from the dataset: empty (NULL), single-character, and a word with perfect translation but a low SSA-COMET-QE score.

et al., 2022), and Google Translate[3].

Results Analysis The benchmark results are presented in Table 5. As the results show, the quality-audited version (corrected MAFAND-MT) outperforms the original evaluation across all settings, indicating that the dataset has been improved. Google Translate outperforms other open-source models, with NLLB-3B being the strongest open alternative in terms of parameter size, followed by NLLB-600M. However, the result remains close to the original. This might be due to two main reasons: 1) the target languages are low-resource languages - the evaluated models do not well represent the languages, 2) the evaluation metrics problem, such as COMET, as discussed in Table 4.

7 Native-Speakers Feedback

Following completion of the translation quality audit, we conducted a survey to gather qualitative insights from native speaker annotators regarding the source text quality, the use of LLM-generated suggestions, and the time they spent on the task. The primary feedbacks are summarized below:

Quality Issues in English Source Text: Annotator feedback revealed significant challenges stemming from the quality and composition of the source (English) text. In particular, source-side noise was observed in segments derived from social media (X/Twitter); as the entries frequently contained platform-specific metadata, such as user handles (@usernames) and hashtags(#), and suffered from syntactic fragmentation due to character limits. Additionally, annotators identified instances of language leakage, where the source text was labeled as English but included content in other languages, which the annotators were unable to interpret. Such issues negatively impacted the annotation process and introduced ambiguity in translation and sentiment interpretation.

Literal Translation: Annotators observed that GPT-5.1 often defaulted to overly literal translations, struggling to balance literal and conceptual meaning, especially for metaphors. This issue was exacerbated by archaic or unnatural terms in human references, which conflicted with modern usage. As a result, current benchmarks may over-reward word-level matching while overlooking native fluency, and, in some cases, GPT-5.1 produced non-existent words in the target language.

8 Conclusions

In this work, we evaluated a subset of a widely used machine translation evaluation dataset for African languages (MAFAND-MT), covering 11 languages and all test set splits, with support from native speakers. The evaluation process involved judging whether each translation was correct or incorrect, labeling the type(s) of translation error(s) for incorrect translations, and producing corrected translations when necessary. Our analysis revealed that the original translations contained various types of errors relative to the source text. We corrected the MAFAND-MT test set using native speakers and LLMs as assistants at different stages. We show that attention should be given to translated evaluation sets, and that relying solely on automatic evaluation metrics for MT quality evaluation may not align with human assessments. A combination of human evaluation, using LLMs as judges, and automatic metrics is recommended. The improved MAFAND-MT test set and the accompanying quality-audit annotation tool, provide valuable resources for researchers conducting further machine translation quality analysis and evaluation.

Limitations

Our work is not without limitations.

First, it focuses on a single MT dataset because recruiting volunteer native speakers for each target language is difficult. However, our pipeline is reproducible and this work can be extended to other African languages' translated dataset such as 1) machine translation dataset: FLORES 101 (Goyal et al., 2022) and FLORES+ (Gordeev et al., 2024) and 2) health (e.g., AFRIDOC-MT (Alabi et al., 2025) and AfriMed-QA (Nimo et al., 2025)), 3)

[3] https://cloud.google.com/translate, Dec 2025

Models	Metrics	amh	hau	ibo	kin	luo	pcm	sna	swa	tsn	twi	yor	Avg.
NLLB 600M	BLEU (MAFAND-MT)	4.92	7.68	17.00	23.11	11.02	7.83	8.72	27.18	25.23	8.10	8.57	13.58
	BLEU (Corrected)	**10.04**	**10.04**	**18.26**	**23.13**	**12.58**	**7.87**	**10.41**	**30.00**	**25.36**	**10.10**	**8.84**	**15.15**
	chrF (MAFAND-MT)	25.49	36.83	47.22	55.70	39.92	27.89	42.55	56.18	56.04	36.87	29.78	41.32
	chrF (Corrected)	**34.06**	**38.81**	**48.10**	**55.72**	**41.61**	**27.93**	**44.56**	**58.32**	**56.09**	**38.43**	**30.22**	**43.08**
Aya 101	BLEU (MAFAND-MT)	3.40	7.12	9.66	09.64	2.84	13.52	5.85	19.96	3.96	2.73	4.22	7.54
	BLEU (Corrected)	**6.10**	**8.98**	**10.27**	**9.61**	**2.92**	**13.50**	**6.32**	**21.83**	**4.02**	**3.25**	**4.21**	**8.27**
	chrF (MAFAND-MT)	20.00	34.88	37.35	37.85	13.94	45.71	27.98	47.77	23.06	23.96	17.16	29.97
	chrF (Corrected)	**25.65**	**36.46**	**37.90**	**37.88**	**14.05**	**45.74**	**28.81**	**49.29**	**23.08**	**24.21**	**17.21**	**30.93**
NLLB 3B	BLEU (MAFAND-MT)	5.62	8.44	20.24	26.60	12.43	4.59	9.39	28.81	28.00	8.69	10.88	14.88
	BLEU (Corrected)	**11.23**	**10.59**	**21.60**	**26.61**	**14.25**	**04.51**	**10.92**	**32.01**	**28.17**	**10.71**	**11.26**	**16.53**
	chrF (MAFAND-MT)	26.51	37.87	50.13	59.28	41.80	**11.63**	43.18	57.54	57.79	38.92	32.36	41.55
	chrF (Corrected)	**35.20**	**39.81**	**51.01**	**59.31**	**43.66**	11.56	**45.02**	**59.91**	**57.85**	**40.71**	**32.83**	**43.35**
Google Trans.	BLEU (MAFAND-MT)	6.35	8.71	15.60	25.33	8.49	00.00	10.69	30.81	31.98	8.87	14.26	14.64
	BLEU (Corrected)	**15.62**	**11.41**	**16.77**	**25.20**	**13.16**	00.00	**12.58**	**34.73**	**32.05**	**10.85**	**14.95**	**17.03**
	chrF (MAFAND-MT)	27.99	38.85	**48.91**	**64.71**	38.02	00.00	45.44	59.52	61.38	40.37	37.33	42.05
	chrF (Corrected)	**39.65**	**41.12**	49.86	64.67	**41.21**	00.00	**47.55**	**62.31**	**61.40**	**42.30**	**37.99**	**44.37**

Table 5: **Zero-shot evaluation benchmark results**. The result compares the original (**MAFAND-MT**) with the **corrected** translation version of the MAFAND-MT test set. All translation directions are from English to target languages. Nigerian Pidgin (pcm) is not supported by Google Translate.

mathematics word problem (e.g., AfriGSM (Adelani et al., 2025b)), and 4) general knowledge and reasoning (e.g., AfriMMLU and AfriXNLI (Adelani et al., 2025b)) and MAFAND-MT dataset (Adelani et al., 2022).

Second, we focused only on the quality audit of the test set, as it is urgent, and research work reports are based on test set results. The same way can be extended for other split sets, such as training and validation sets.

References

Idris Abdulmumin, Michael Beukman, Jesujoba O. Alabi, Chris Emezue, Everlyn Asiko, Tosin Adewumi, Shamsuddeen Hassan Muhammad, Mofetoluwa Adeyemi, Oreen Yousuf, Sahib Singh, and Tajuddeen Rabiu Gwadabe. 2022. Separating Grains from the Chaff: Using Data Filtering to Improve Multilingual Translation for Low-Resourced African Languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1001–1014, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Idris Abdulmumin, Sthembiso Mkhwanazi, Mahlatse Mbooi, Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Neo Putini, Miehleketo Mathebula, Matimba Shingange, Tajuddeen Gwadabe, and Vukosi Marivate. 2024. Correcting FLORES Evaluation Dataset for Four African Languages. In *Proceedings of the Ninth Conference on Machine Translation*, pages 570–578, Miami, Florida, USA. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen H. Muhammad, Guyo D. Jarso, Oreen Yousuf, and 26 others. 2022. A Few Thousand Translations Go a Long Way! Leveraging Pre-trained Models for African News Translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, and 1 others. 2025a. 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.

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. 2025b. 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 Oluwadara Alabi, Israel Abebe Azime, Miaoran Zhang, Cristina España-Bonet, Rachel Bawden, Dawei Zhu, David Ifeoluwa Adelani,

Clement Oyeleke Odoje, Idris Akinade, Iffat Maab, Davis David, Shamsuddeen Hassan Muhammad, Neo Putini, David O. Ademuyiwa, Andrew Caines, and Dietrich Klakow. 2025. AFRIDOC-MT: Document-level MT Corpus for African Languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27758–27794, Suzhou, China. Association for Computational Linguistics.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the Translation Initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Tadesse Destaw Belay, Kedir Yassin Hussen, Sukairaj Hafiz Imam, Ibrahim Said Ahmad, Isa Inuwa-Dutse, Abrham Belete Haile, Grigori Sidorov, Iqra Ameer, Idris Abdulmumin, Tajuddeen Gwadabe, Vukosi Marivate, Seid Muhie Yimam, and Shamsuddeen Hassan Muhammad. 2025. The Rise of AfricaNLP: Contributions, Contributors, and Community Impact (2005-2025). *Preprint*, arXiv:2509.25477.

Nathaniel Berger, Stefan Riezler, Miriam Exel, and Matthias Huck. 2024. Prompting Large Language Models with Human Error Markings for Self-Correcting Machine Translation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 636–646, Sheffield, UK. European Association for Machine Translation (EAMT).

Júlia Falcão, Claudia Borg, Nora Aranberri, and Kurt Abela. 2024. COMET for low-resource machine translation evaluation: A case study of English-Maltese and Spanish-Basque. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3553–3565, Torino, Italia. ELRA and ICCL.

Zhaopeng Feng, Yan Zhang, Hao Li, Bei Wu, Jiayu Liao, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2025a. TEaR: Improving LLM-based Machine Translation with Systematic Self-Refinement. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3922–3938, Albuquerque, New Mexico. Association for Computational Linguistics.

Zhaopeng Feng, Yan Zhang, Hao Li, Bei Wu, Jiayu Liao, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2025b. TEaR: Improving LLM-based machine translation with systematic self-refinement. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3922–3938, Albuquerque, New Mexico. Association for Computational Linguistics.

Freitag, Markus and Foster, George and Grangier, David and Ratnakar, Viresh and Tan, Qijun and Macherey, Wolfgang. 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.

Isai Gordeev, Sergey Kuldin, and David Dale. 2024. FLORES+ Translation and Machine Translation Evaluation for the Erzya Language. In *Proceedings of the Ninth Conference on Machine Translation*, pages 614–623, Miami, Florida, USA. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A Survey on LLM-as-a-Judge. *Preprint*, arXiv:2411.15594.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES Evaluation Datasets for Low-Resource Machine Translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Lifeng Han, Alan Smeaton, and Gareth Jones. 2021. Translation Quality Assessment: A Brief Survey on Manual and Automatic Methods. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 15–33, online. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Shaoxiong Ji, Zihao Li, Jaakko Paavola, Hengyu Luo, and Jörg Tiedemann. 2025. Massively Multilingual

Adaptation of Large Language Models Using Bilingual Translation Data. *arXiv preprint 2506.00469*.

Ahrii Kim. 2025. RUBRIC-MQM : Span-Level LLM-as-judge in Machine Translation For High-End Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 147–165, Vienna, Austria. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, and 33 others. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Senyu Li, Jiayi Wang, Felermino D. M. A. Ali, Colin Cherry, Daniel Deutsch, Eleftheria Briakou, Rui Sousa-Silva, Henrique Lopes Cardoso, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. SSA-COMET: Do LLMs Outperform Learned Metrics in Evaluating MT for Under-Resourced African Languages? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12990–13009, Suzhou, China. Association for Computational Linguistics.

Jessy Lin, Geza Kovacs, Aditya Shastry, Joern Wuebker, and John DeNero. 2022. Automatic Correction of Human Translations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 494–507, Seattle, United States. Association for Computational Linguistics.

Yixin Liu, Kejian Shi, Katherine He, Longtian Ye, Alexander Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2024. On Learning to Summarize with Large Language Models as References. 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 8647–8664, Mexico City, Mexico. Association for Computational Linguistics.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024. Error Analysis Prompting Enables Human-Like Translation Evaluation in Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8801–8816, Bangkok, Thailand. Association for Computational Linguistics.

Charles Nimo, Tobi Olatunji, Abraham Toluwase Owodunni, Tassallah Abdullahi, Emmanuel Ayodele, Mardhiyah Sanni, Ezinwanne C. Aka, Folafunmi Omofoye, Foutse Yuehgoh, Timothy Faniran, Bonaventure F. P. Dossou, Moshood O. Yekini, Jonas Kemp, Katherine A Heller, Jude Chidubem Omeke, Chidi Asuzu Md, Naome A Etori, Aïmérou Ndiaye, Ifeoma Okoh, and 7 others. 2025. AfriMed-QA: A Pan-African, Multi-Specialty, Medical Question-Answering Benchmark Dataset. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1948–1973, Vienna, Austria. Association for Computational Linguistics.

Antoine Nzeyimana. 2024. Low-resource neural machine translation with morphological modeling. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 182–195, Mexico City, Mexico. Association for Computational Linguistics.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. AfroBench: How Good are Large Language Models on African Languages? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19048–19095, Vienna, Austria. Association for Computational Linguistics.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b & gpt-oss-20b Model Card. *Preprint*, arXiv:2508.10925.

OpenAI, Albert : Jiang, Alexandre Sablayrolles, Alexis Tacnet, Alok Kothari, Antoine Roux, Arthur Mensch, Audrey Herblin-Stoop, Augustin Garreau, Austin Birky, Bam4d, Baptiste Bout, Baudouin de Monicault, Blanche Savary, Carole Rambaud, Caroline Feldman, Devendra Singh Chaplot, Diego de las Casas, Diogo Costa, and 51 others. 2024. Mistral Large 2. `https://huggingface.co/mistralai/Mistral-Large-Instruct-2407`. Large-scale instruct-tuned language model released by Mistral AI.

Shenbin Qian, Archchana Sindhujan, Minnie Kabra, Diptesh Kanojia, Constantin Orasan, Tharindu Ranasinghe, and Fred Blain. 2024. What do Large Lan-

guage Models Need for Machine Translation Evaluation? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3660–3674, Miami, Florida, USA. 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.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Pratik Rakesh Singh, Kritarth Prasad, Mohammadi Zaki, and Pankaj Wasnik. 2025. In-Domain African Languages Translation Using LLMs and Multi-armed Bandits. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 167–175, Vienna, Austria. Association for Computational Linguistics.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large Language Models for Data Annotation and Synthesis: A Survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

Gongbo Tang, Oreen Yousuf, and Zeying Jin. 2024. Improving BERTScore for Machine Translation Evaluation Through Contrastive Learning. *IEEE Access*, 12:77739–77749.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 Technical Report. *Preprint*, arXiv:2503.19786.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *Preprint*, arXiv:2207.04672.

Atnafu Lambebo Tonja, Israel Abebe Azime, Tadesse Destaw Belay, Mesay Gemeda Yigezu, Moges Ahmed Ah Mehamed, Abinew Ali Ayele, Ebrahim Chekol Jibril, Michael Melese Woldeyohannis, Olga Kolesnikova, Philipp Slusallek, Dietrich Klakow, and Seid Muhie Yimam. 2024. EthioLLM: Multilingual Large Language Models for Ethiopian Languages with Task Evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6341–6352, Torino, Italia. ELRA and ICCL.

Marcos V Treviso, Nuno M Guerreiro, Sweta Agrawal, Ricardo Rei, José Pombal, Tania Vaz, Helena Wu, Beatriz Silva, Daan Van Stigt, and Andre Martins. 2024. xTower: A Multilingual LLM for Explaining and Correcting Translation Errors. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15222–15239, Miami, Florida, USA. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Pavanpankaj Vegi, Sivabhavani J, Biswajit Paul, Prasanna K R, and Chitra Viswanathan. 2022a. ANVITA-African: A Multilingual Neural Machine Translation System for African Languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1090–1097, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Pavanpankaj Vegi, Sivabhavani J, Biswajit Paul, Abhinav Mishra, Prashant Banjare, Prasanna K R, and Chitra Viswanathan. 2022b. WebCrawl African : A Multilingual Parallel Corpora for African Languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1076–1089, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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. 2024. 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.

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

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Appendix

A Translation Annotation Guideline

The native speakers do not have information about the two transition option sources, which are the original human translation and LLM translation. The native speakers were given the following guidelines.

Reviewing translation errors: At this stage, the native speakers read and evaluate the translation quality in parallel with the given source English text.

- Read the source English sentence.
- Read the translated in Translation 1 and Translation 2 (One is the original translation, and the other is the LLM translation; we randomly shuffle the content positions of the two translations when displaying for the annotators).
- For both Translation 1 and Translation 2, choose Correct or Incorrect.
- If any of the translations are `Incorrect`, select one or more error types (multi-label error selection approach) that best describe the error by ticking from the list of error types, shown in Table 2.

Correcting the translations: If both Translation 1 and Translation 2 are marked Incorrect with the corresponding error types, provide a new correct translation in the given text box; the UI is shown in Appendix B.

B Annotation Tool UI

Figure B shows a screenshot of our machine translation quality audit annotation tool UI.

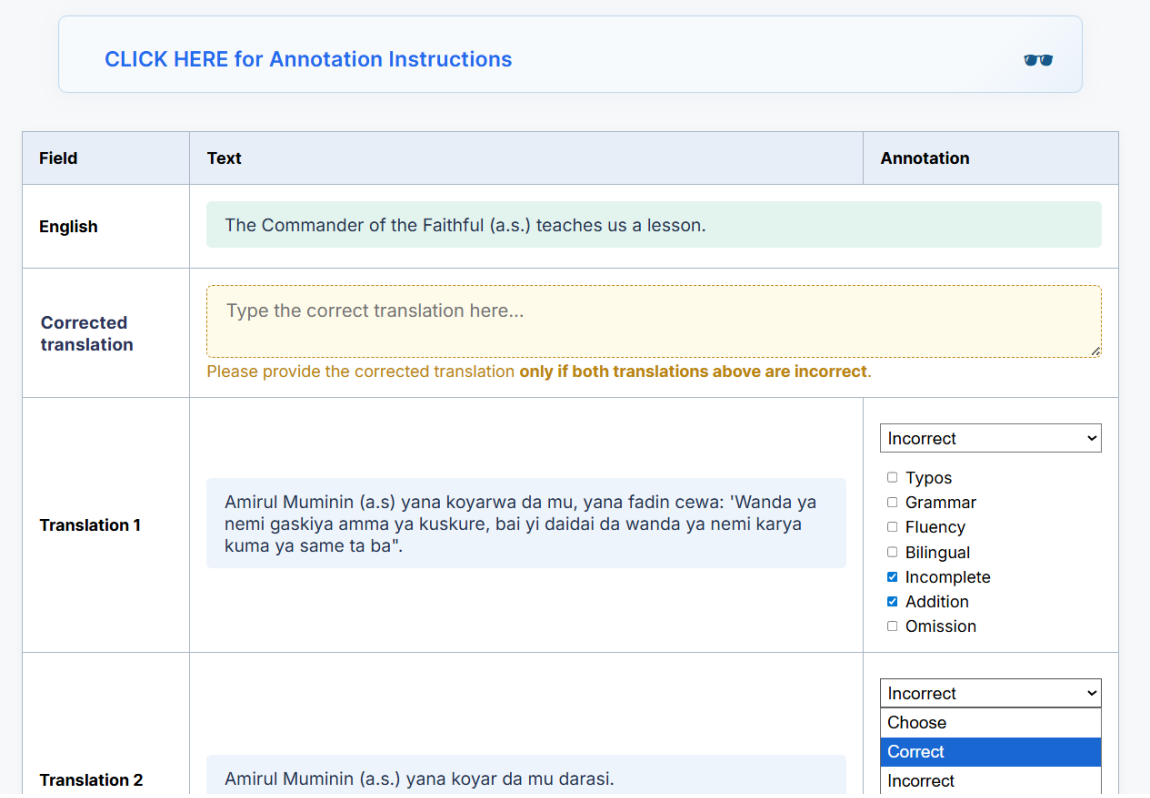

Figure 5: **Annotation tool UI for Hausa language**. The tool will be publicly released upon acceptance of the work for further machine translation and other NLP dataset quality audits with additional features.

C LLM-as-a-Judge Prompts

Prompt: LLM-as-a-judge for translation quality analysis

```
You are an expert translation quality analyst with deep knowledge of machine
translation evaluation from English to African languages. Analyze the following
English → {target_lang_name} translation..


 Possible translation error types (choose one or more when incorrect):
- Typos : misspellings or character mistakes in the translation
- Grammar : grammatical errors (agreement, tense, syntax)
- Fluency : unnatural or awkward phrasing / non-native flow
- Bilingual : interference or literal translation from English
- Incomplete : translation omits part(s) of the source meaning
- Addition : adds information not present in the source
- Omission : removes information present in the source


SOURCE (English):
{eng_text}
TRANSLATION ({target_language_name}):
{tgt_text}

Your task is to follow the below rules exactly:
1) Decide if the translation is correct or incorrect contextually. If correct,
respond with status "correct". Only mark as 'incorrect' when the meaning changes.
Do NOT mark minor differences as errors.
2) If incorrect, set status "incorrect", pick one or more error types from the
taxonomy, and give a short explanation for each type of error.
3) IF incorrect, ALSO PROVIDE a corrected translation in the target language in
the field "correct_translation" (a fluent, natural translation that preserves
source meaning correctly).
4) If correct, set "correct_translation" to null.
5) Return ONLY valid JSON (no extra commentary). Use this exact structure:


{{
"status": "correct" | "incorrect",
"errors": [
{{"type": "<one_of_taxonomy>", "description": "<short explanation>"}}
],
"correct_translation": "<correct text or null>"
}}
Taxonomy reference:
{taxonomy} Return only the JSON.
```

D Translation Error Labeling Agreement Between Human vs LLMs

Figure 6 the overlap between human and LLM-as-a-judge for translation error labeling. The statistics show that LLMs are overpredicting error types relative to humans in the targeted low-resource languages.

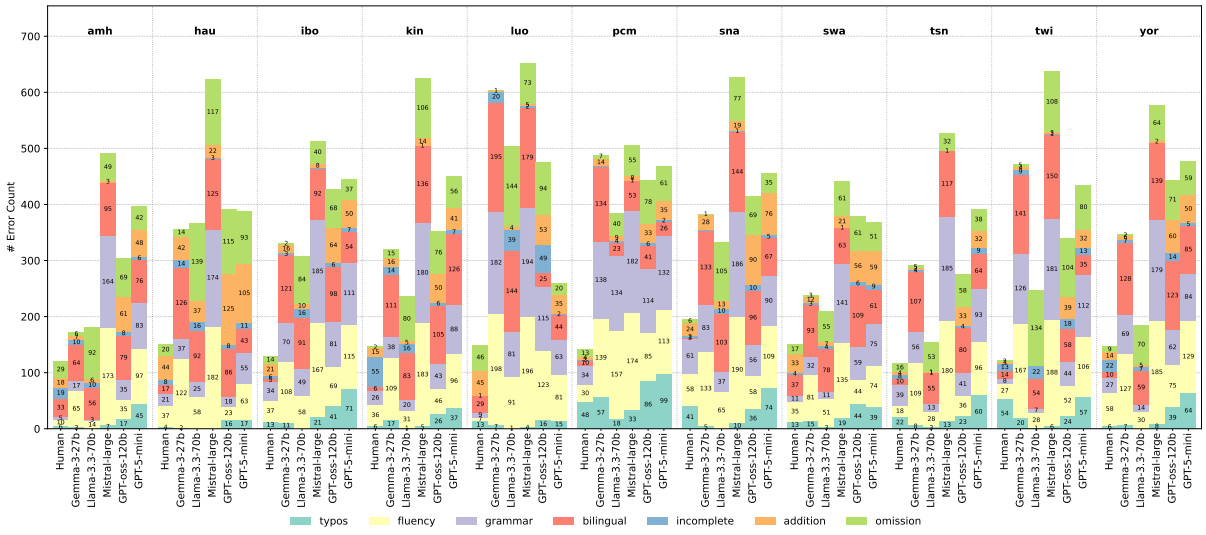

Figure 6: **Translation error labeling overlap between Human and LLM-as-a-judge.**