Trust but Check: LLM-Assisted Quality Control for Human Translated Benchmarks

Anonymous ACL submission

Abstract

Machine translation (MT) benchmarks for low-resource languages are often built from human translations of high-resource source data (e.g., English, French). Even professionally produced references contain adequacy and fluency errors that propagate into model rankings. Manual auditing at scale is prohibitively expensive, and fully automatic correction risks encoding model-specific biases into the benchmark. We study a *trust-but-verify* pipeline in which an LLM flags potential translation errors and native speakers adjudicate and correct the flagged cases. We apply this to the MAFAND-MT test set across 11 African languages, where the LLM flags 58% of references as incorrect. Evaluating classic MT systems on the human-verified references gives improvements in BLEU, chrF, and COMET. We also analyze COMET on low-resource languages and find that scores remain modest even for exact reference matches. We release the human-verified MAFAND-MT references to support MT evaluation for 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 (amh), Kinyarwanda (kin), and Hausa (hau) languages contains such errors, creating misconceptions during MT model evaluation (Liu et al., 2024b). 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, such as MT error correction (Treviso et al., 2024; Feng et al., 2025a). Most available evaluation datasets for low-resource languages are created by translating existing datasets by native speakers. For high-resource languages such as English, LLMs are assisting in dataset creation and annotation , a role for which LLMs are often ill-suited for low-resource

English Source	MAFAND-MT Translation	Correct Translation	Trans. Errors
"eng": "Date: Thursday, July 31, 2014"	"amh":'የዘን ዘጠኝ ጦማሪዎች ፍትህ ይገበቸዋል"	"amh": ቀን፡ሐሙስ፣ ሐምሌ 31፣ 2014	Incomplete, Omission
"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 ..."	"kin": "Ian Simbota yari ahagarariye ishyirahamwe ry'abantu bafite ubumuga bw'uruhu mu rukiko."	Typos, Addition, Omission
"eng": "Every soul shall have a taste of death."	"hau": "Ko shakka babu akwai wani lokaci da rayuwa za ta zo karshe "Kowane rai mai dandanar mutuwa ne" (Suratu Al Imrana 3:185)."	"hau": "Ubangiji, muna d⌒aukin jiran wannan damar"	Addition, Fluency

Table 1: **Examples of translation errors from the MAFAND-MT test dataset** (Adelani et al., 2022). The **MAFAND-MT Translation** column is the incorrect translation of the given English-sourced (eng) text, and the **Correct Translation** is the correct translation using native speakers of the target language.

languages due to translation quality and cultural grounding. However, how current LLMs assist in refining translated datasets for low-resource languages remains unexplored. In this work, we investigate whether LLMs can act as *first-pass auditors*, surfacing likely errors for expert review rather than replacing human judgment. We audit an existing evaluation benchmark and release human-verified references only where native speakers accept or override LLM suggestions.

Specifically, we ask: **RQ1** Can LLMs reliably detect incorrect human references and propose useful corrections for low-resource languages? **RQ2** What types of translation errors are present in the professionally translated references of the MAFAND-MT dataset, and how prevalent are they? **RQ3** How do human-verified references affect MT leaderboard scores across model families and under self-preference evaluation?

Our contributions are: (1) an LLM-assisted, human-verified auditing framework for refining low-resource MT benchmarks; (2) a multi-label error analysis of human translation references, including LLM–human agreement on the error taxonomy; (3) an LLM-audited and human-corrected version of the MAFAND-MT dataset; and (4) benchmark results demonstrating consistent improvements across diverse MT systems and evaluation metrics, including chrF++, BLEU, and COMET. Together, these findings position LLM auditing as a scalable *triage* mechanism rather than an autonomous solution for dataset repair.

2 Related Work

2.1 Auditing Training Corpora Quality

Early work efforts 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 to improve NLP systems. However, the reliance on web-crawled data introduced substantial noise, including mistranslations, misalignment, 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. We position LLMs as a complementary tool for MT data auditing, examining their ability to judge whether the given translation is correct or not, identify translation errors, and suggest correct translations.

2.2 Auditing Test Dataset Quality

Language models are commonly evaluated on downstream tasks, with MT serving as a central benchmark for assessing cross-lingual capabilities. For African languages, MAFAND-MT (Adelani et al., 2022) is the 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 of 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 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. In this work, we explore 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 for 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., 2024a). Within MT research, LLM-assisted translation error detection and correction have been explored primarily in English for 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 translation correction (Lin et al., 2022).

Despite these advances, the use of LLMs as translation reviewers, capable of identifying, categorizing, and correcting translation errors, remains unexplored for low-resource 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. Specifically, we assess their ability to judge whether the translation is *correct* or *incorrect*, identify common *types of translation errors*, propose *correction candidates*, and support a human-in-the-loop auditing pipeline. By analyzing 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. For machine translation NLP tasks, prominent datasets that include African languages are translated from high-resource language sources, such as 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 translated by native speakers of the target languages (Adelani et al., 2022). However, we observed a range of common MT error types, as shown in Table 1. We review the translation quality using LLMs as a helper and categorize the error types into seven MT error types (typos, grammar, fluency, bilingual, incomplete, addition, and omission) adopted from previous work (Liu et al., 2024b; Kim, 2025).

4 Translation Review Pipeline

We used a two-stage review pipeline to assess and review translation quality. In the first stage, LLMs automatically evaluate each original translation pair, flag potential errors, and suggest correct translations if the pair is flagged as incorrect. In the

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

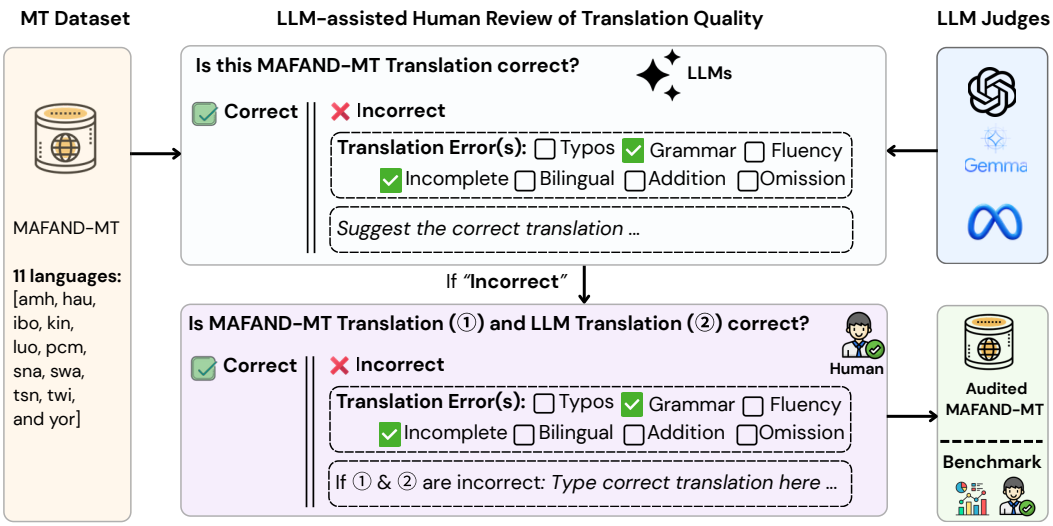

Figure 1: LLM-assisted auditing pipeline for MT dataset quality assessment and correction. The figure illustrates the workflow for judging whether the translation is correct or not, identifying translation errors, and generating correction suggestions.

second stage, native speakers verify the flagged cases and, where necessary, provide a new translation if there is mistranslation in both the original and the LLM-suggested translation, as shown in Figure 1.

4.1 LLM-assisted Translation Review

We select the following popular LLMs for the LLM-as-a-judge evaluation: GPT-5.1 [2] and Gemini-3.1-Pro[3], Gemma-3-27B (Kamath et al., 2025), GPT-oss-120B (OpenAI et al., 2025), Gemma-3-27B (Team et al., 2025), Llama-3.3-70B (Grattafiori et al., 2024), Mistral-123B (OpenAI et al., 2024), and Qwen3.5-27B (Yang et al., 2025) as a primary translation auditor.

Selecting better LLM To assess all responses of LLM outputs using humans is costly and time-consuming; we select a better LLM based on human-LLM agreement for end-to-end assistance. We first select 100 random samples for each language and have them audited by humans to determine whether they are correct/incorrect translations and, if incorrect, the types of errors. We then compare the LLM's responses to the same prompts with human responses. GPT-5.1 has better agreement with humans, leading with 58% agreement with correct/incorrect labeling and 27% with the multi-label error type agreement. In additon, while open-weights models like Gemma 3 27B and Qwen3.5 27B exhibit severe negative-class prediction biases on the low resource languages, and Gemini 3.1 suffers from systemic all-or-nothing failures (e.g., 0% negative detection in Hausa while there are mentioned some real examples in the motivation table, Table 1, 100% in Luo), GPT 5.1 maintains a robust, active distribution across all test sets, making it the better reliable predictor comparably to reduce human validation fatigue without introducing catastrophic model bias. The statistics of LLMs on the predictions of correct/incorrect are shown in Table 2.

4.2 Human Post-Edits

Native speaker selection We recruited at least two native-speaker volunteers per language (11 languages total), coordinated through community groups and co-author networks. Annotators were native in the target language. We select native speakers for each targeted language with the criterion that at least one publication in the NLP research area to understand the source English text and the task itself.

The role of native speakers Native speakers adjudicate every GPT-flagged item: they label both translation options as correct or incorrect, select error types, and write a new reference when both given options (the MAFAND-MT and the LLM-suggested translation are provided as options for the native speakers) fail. To facilitate the auditing task for native speakers, we design an interac-

[2] https://openai.com/index/gpt-5-1/
[3] https://deepmind.google/models/gemini/pro/

			GPT 5.1		Gemini 3.1		Gemma 3 27 B		Qwen3.5 27B		LLama 3 70B	
Lang.	**Trans.**	**# Test set**	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗
amh	`eng-amh`	1,037	271	766	174	833	210	826	77	960	214	823
hau	`eng-hau`	1,500	680	820	1500	0	176	1324	65	1435	284	1216
ibo	`eng-ibo`	1,500	884	964	627	870	298	1202	242	1258	519	981
kin	`eng-kin`	1,006	390	616	172	832	367	639	132	874	368	638
luo	`eng-luo`	1,500	398	1,102	1500	0	0	1500	1	1499	17	1483
pcm	`eng-pcm`	1,564	1009	555	490	1070	278	1286	89	1475	21	1543
sna	`eng-sna`	1,005	368	637	170	826	84	921	39	966	103	902
swa	`eng-swa`	1,835	748	787	583	1240	733	1102	378	1457	986	849
tsn	`eng-tsn`	1,500	844	656	466	1027	158	1342	78	1422	790	710
twi	`eng-twi`	1,500	265	1285	97	1401	12	1488	7	1493	134	1366
yor	`eng-yor`	1,558	585	973	254	1277	112	1446	65	1493	750	808

Table 2: The MAFAND-MT test set dataset details: languages, translation direction, number of test set instances, the statistics of corresponding LLM says **correct** (✓) and **incorrect** (✗) from a given **Test set**. Other LLMs, such as GPT-oss-120b and Mistral-large, exhibit a strong initial tendency to predict all *incorrect*.

tive 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 provided options are incorrect.

Human-in-the-Loop Auditing Tool We use the LLM-suggested feedback as a helper for native speaker quality inspections by designing a new interactive annotation tool that presents the source (English) text, two translation versions (original translation and LLM-suggested translation), common types of translation errors if any of the translations are incorrect, and a text field to write new translation version if both of the provided translations are incorrect. The auditing tool is customized from the POTATO annotation tool (Pei et al., 2022), and the details of the translation review guidelines with the annotation interface are provided in Appendix A.

5 The Role of LLMs for MT Auditing

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

Can LLMs serve as judges of translation quality for low-resource languages? Based on the agreement between GPT-5.1 and humans, GPT-5.1 pre-screening is used as triage; native speakers adjudicate only the 9,161 pairs GPT flagged (58.7% of the test set); the remaining 41.29% were auto-

Outcome	Count	% reviewed
Both translations acceptable	851	12.3
Human rejects GPT (MAFAND acceptable)	1,883	27.2
Humans prefer GPT (MAFAND faulty)	1,942	28.0
Both faulty (new reference written)	2,252	32.5

Table 3: **Human adjudication on GPT-flagged items (n=6,928 reviewed pairs).** Annotators saw the MAFAND reference and a GPT-5.1 alternative. Released `corrected` references were updated only in the both *incorrect* row.

passed when GPT labeled the MAFAND reference `correct`, to ensure that every GPT-`incorrect` pair is human-reviewed. When humans also rejected the original MAFAND text, they agreed with GPT's `incorrect` verdict in 60.5% of reviewed items (precision 0.61 for GPT-as-detector). Other LLMs mostly say *incorrect* based on the agreement of 200 samples after the final human-audited instance, as shown in Figure 2.

Table 3 summarizes patterns between human and GPT-5.1 during auditing. Annotators preferred the GPT alternative over the MAFAND reference in 28.0% of reviewed cases (1,942), and rated both options acceptable in 12.3%. Critically, **none** of the 1,942 "prefer GPT" judgments were copied into the released `corrected` field. Conversely, **99.96%** of the 2,251 benchmark edits (14.5% of the full test set) came from the 32.5% of items where *both* translations were rejected, and annotators wrote a third new reference. LLM suggestions, therefore, informed the reviewer's opinion but did not automatically flow into the benchmark; only explicit human rewrites did. Therefore, LLMs help

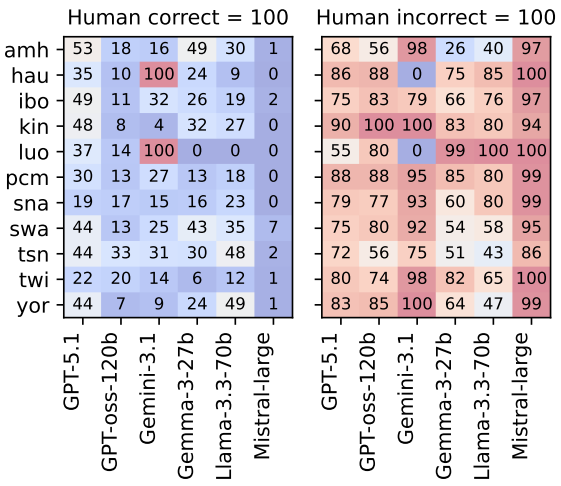

Figure 2: Human judgments versus LLM-as-a-judge on 200 randomly selected samples: 100 translations labeled as `correct` (left bar graph) and 100 as `incorrect` (right bar graph) by human evaluators.

	Human	**GPT-5.1**
Mean error types per sentence	1.47	2.39
Sentences with ≥ 3 error types	9.5%	45.5%
Incomplete among dual-reject cases	10%	59%
Mean Jaccard overlap of error sets	0.27	

Table 4: Error-labeling behavior when human and GPT both mark the MAFAND reference incorrect (n=4,194). GPT assigns more labels per sentence and favors *Incomplete* omissions.

humans by suggesting a better translation in 28% of instances of incorrect translation.

Can LLMs be in a translation error-type annotator for low-resource languages? Table 4 characterizes error typing when both raters reject the MAFAND reference. GPT assigns 2.39 error types per sentence versus 1.47 for humans, and applies three or more labels 45.5% of the time (9.5% for humans). The label *Incomplete* appears in 59% of GPT diagnoses but only 10% of human diagnoses on the same sentences (mean Jaccard overlap 0.27). When humans rejected GPT while accepting MAFAND (n=1,883), the dominant critiques of GPT text were fluency and grammar, not typos. LLMs (specifically GPT-5.1) classify most things incorrectly, over-predicting error-type labels, even though native speakers say it correctly, and classify errors more than native speakers, as shown in the summary table, Table 4.

How prevalent are translation errors in existing low-resource machine translation benchmark datasets? Figure 3 shows human-labeled error types from the targeted MAFAND-MT dataset.

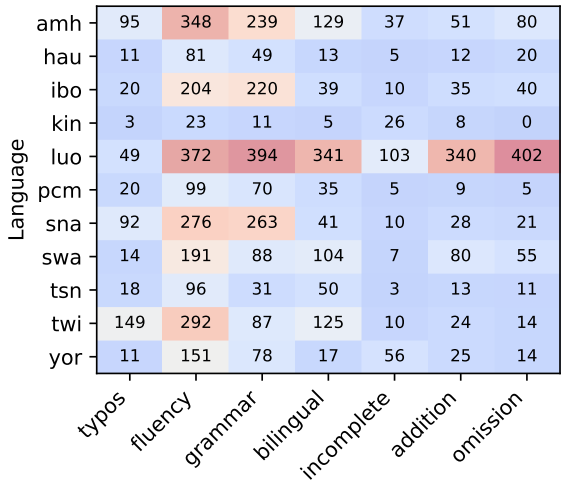

Figure 3: The statistics of Human-labeled translation error counts for the full MAFAND-MT test dataset.

The error analysis of the MAFAND-MT benchmark shows that translation errors are highly prevalent across most evaluated low-resource African languages, indicating significant quality issues in existing MT benchmark datasets. Fluency and grammar errors dominate, with particularly high frequencies in languages such as Luo, Amharic, Twi, and Shona; for instance, Luo exhibits large numbers of grammar, fluency, bilingual interference, and omission errors, suggesting unnatural phrasing and missing or distorted content. Similar patterns appear in Amharic and Shona, while Twi shows notable typo-related issues. Overall, these results show that current low-resource MT benchmarks contain systematic translation problems that can bias evaluation results, distort leaderboard rankings, and undermine reliable MT comparison. The presence of such a frequency of human-identified translation errors can significantly affect MT evaluation systems in the absence of proper auditing.

6 Benchmarking Audited MAFAND-MT

Automated metrics provide a cost-effective proxy for translation quality and are widely used to track MT system progress (Kocmi et al., 2024). We assess the impact of benchmark auditing along three dimensions: (1) native-speaker judgments collected during the auditing process, (2) automatic evaluation metrics including BLEU, chrF++, and SSA-COMET-QE computed on the reference translations, and (3) end-to-end MT performance across diverse systems, including traditional MT models

Lang.	NLLB 600M				NLLB 3B				Aya 101				Google Translate			
	BLEU		chrF		BLEU		chrF		BLEU		chrF		BLEU		chrF	
	MTv1	MTv2	MTv1	MTv2	MTv1	MTv2	MTv1	MTv2	MTv1	MTv2	MTv1	MTv2	MTv1	MTv2	MTv1	MTv2
amh	4.92	**10.04**	25.49	**34.06**	3.40	**6.10**	20.00	**25.65**	5.62	**11.23**	26.51	**35.20**	6.35	**15.62**	27.99	**39.65**
hau	7.68	**10.04**	36.83	**38.81**	7.12	**8.98**	34.88	**36.46**	8.44	**10.59**	37.87	**39.81**	8.71	**11.41**	38.85	**41.12**
ibo	17.00	**18.26**	47.22	**48.10**	9.66	**10.27**	37.35	**37.90**	20.24	**21.60**	50.13	**51.01**	15.60	**16.77**	48.91	**49.86**
kin	23.11	**23.13**	55.70	**55.72**	9.64	**9.61**	37.85	**37.88**	26.60	**26.61**	59.28	**59.31**	25.33	25.20	64.71	64.67
luo	11.02	**12.58**	39.92	**41.61**	2.84	**2.92**	13.94	**14.05**	12.43	**14.25**	41.80	**43.66**	8.49	**13.16**	38.02	**41.21**
pcm	7.83	**7.87**	27.89	**27.93**	13.52	13.50	45.71	**45.74**	4.59	4.51	11.63	11.56	0.00	0.00	0.00	0.00
sna	8.72	**10.41**	42.55	**44.56**	5.85	**6.32**	27.98	**28.81**	9.39	**10.92**	43.18	**45.02**	10.69	**12.58**	45.44	**47.55**
swa	27.18	**30.00**	56.18	**58.32**	19.96	**21.83**	47.77	**49.29**	28.81	**32.01**	57.54	**59.91**	30.81	**34.73**	59.52	**62.31**
tsn	25.23	**25.36**	56.04	**56.09**	3.96	**4.02**	23.06	**23.08**	28.00	**28.17**	57.79	**57.85**	31.98	**32.05**	61.38	**61.40**
twi	8.10	**10.10**	36.87	**38.43**	2.73	**3.25**	23.96	**24.21**	8.69	**10.71**	38.92	**40.71**	8.87	**10.85**	40.37	**42.30**
yor	8.57	**8.84**	29.78	**30.22**	4.22	4.21	17.16	**17.21**	10.88	**11.26**	32.36	**32.83**	14.26	**14.95**	37.33	**37.99**

Table 5: **Zero-shot evaluation benchmark results**. The result compares the original (**MAFAND-MT - MTv1**) with the **corrected - MTv2** (human-audited) 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. BLEU and chrF are imported from sacrebleu.

(NLLB-600M, NLLB-3B), instruction-tuned multilingual LLMs (Aya-101), and commercial MT systems such as Google Translate. All automatic evaluation scores are reported using both the original and the human-finalized references. For bias analysis, we additionally compare results against the unaudited GPT-5.1 correction outputs extracted from the audit logs.

Results analysis Table 5 reports BLEU and chrF for NLLB-600M, NLLB-3B, Aya-101, and Google Translate against original and the new human-audited references. Every system gains under the audited references (BLEU +0.4 to +9.3; chrF +0.3 to +8.7; SSA-COMET-QE +0.05 to +3.81), confirming that the benchmark was systematically noisy rather than favoring a single model family. Based on the total correct and incorrect judgments from the LLM shown in Table 2, and the overall proportions of translation error types identified by LLMs compared to human auditors in Figure 2, we find that relying on LLM judgments without a human-in-the-loop remains unsafe for low-resource languages. LLMs can still effectively assist human auditors by providing an additional option for evaluation rather than fully replacing human translation or annotation from scratch. The improvement across all evaluation metrics, such as chrF, BLEU, and SSA-COMET-QE score, indicates that the dataset is of higher quality. However, just because a system is perceived to do better, does not mean that the data used to assess it is, itself, a better dataset. In the following section, we further explore the evaluation metrics, such as SSA-COMET-QE, as the improvement in results is not very significant and is contextual (captures the relationship between the source and target languages, rather than character/word occurrence).

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 6: 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.

Is the COMET score a reliable quality estimation for low-resource languages? SSA-COMET-QE (Sub-Saharan African Cross-lingual Optimized Metric for Evaluation of Translation) - is an improved version of AfriCOMET (Wang et al., 2024), a robust and automatic metric for machine translation quality estimation (Rei et al., 2020). The score ranges from 0 (semantically unrelated) to 1 (high quality) that reflects the quality of the translation (Li et al., 2025). The SSA-COMET-QE quality estimation score for the MAFAND-MT dataset before and after quality check is presented in Appendix Figure 6. There is a statistical improvement from the lower threshold (<0.65) to the higher threshold (>0.65) after auditing, the increment ranging from 6 to 172 transaction pairs per language. The quality estimation score is presented in Table 7.

However, as illustrated in Table 6, an empty source and output, or a single-character or single-word translation, can still receive a low score. Across all languages, the maximum score observed is 0.84; even perfect translations do not approach

Metric	amh	hau	ibo	kin	luo	pcm	sna	swa	tsn	twi	yor	Avg.
MAFAND-MT (MTv1)	60.99	66.91	65.49	67.03	69.31	44.77	67.39	67.58	65.23	63.84	64.38	63.90
Corrected (MTv2)	**64.50**	**67.09**	**65.68**	**67.15**	**69.69**	**44.85**	**67.87**	**67.98**	**65.29**	**65.03**	**64.47**	**64.51**

Table 7: COMET evaluation scores using original MAFAND-MT (MTv1) and corrected (MTv2) references. Higher scores indicate better translation quality. COMET is SSA-COMET-QE and calculated with our reference using the source, and MAFAND-MT or Corrected, accordingly

.

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 behavior of these evaluation metrics in such settings. While we improved the statistics for low-scoring translations, the improvement remains modest given the quality of the SSA-COMET-QE metrics, as shown in Table 6.

Reference bias and LLM self-preference Rescoring under three references: (i) MAFAND original, (ii) unaudited GPT-5.1 corrections, (iii) human-final text, shows Gemini chrF rising from 45.9 (original) to 48.1 (human-final) but to 53.7 against GPT references, while GPT correction strings score 100 chrF on GPT references. Human verification is required so leaderboard gains reflect improved ground truth, not evaluator–generator self-preference.

Human–LLM cost trade-off We conducted a survey to gather qualitative insights from native speaker annotators. Annotators reported 30 s–3 min per item after LLM pre-screening (64.7% at 1–3 min), and 42% found the extra LLM option helpful versus 18% not helpful. Native speakers responded: 42.2% yes, it was helpful; 42.1% partially useful; and 17.6% not helpful. LLM-first triage reduces the human workload to disputed cases but does not remove the need for native-speaker adjudication on accept/reject and final wording.

Quality Issues in the Source Text: A significant challenge stemming from the quality of the source (English) text is that 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. Need to consider the nature of the source text carefully before translating into other languages in the future. Our approach instead uses a better LLM to prioritize sentence pairs most likely to contain problematic translations, so human effort is focused on a much smaller subset (9K in the case of GPT-5.1, which is 58% of the total test set). In effect, the LLM filters the pool from "review everything" to "review the most suspicious items," which directly reduces auditing time and cost.

7 Conclusions

We audited the MAFAND-MT test references across 11 African languages using a GPT-5.1-assisted triage pipeline with native-speaker verification. Our findings reveal that professionally translated benchmark references still contain substantial adequacy and fluency errors. While GPT-5.1 is effective at surfacing potential translation issues, it tends to over-flag errors and shows limited agreement with human annotators, achieving 58% agreement on correct/incorrect judgments and only 25% agreement on multi-label translation error categorization. These results suggest that LLMs can serve as useful assistants for MT quality auditing when combined with human-in-the-loop verification, and that proprietary models currently provide more reliable auditing performance for low-resource languages than open-source LLMs. We further show that human-verified references improve evaluation outcomes across diverse MT systems and automatic metrics, including chrF++, BLEU, and COMET. Based on these findings, we recommend that practitioners treat LLM-based auditing as a cost-aware triage mechanism, explicitly report agreement with human judgments, and evaluate MT systems using human-approved references. To support reproducibility and future benchmark maintenance, we will release the audited MAFAND-MT test set together with the auditing tool.

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. We evaluate only GPT-5.1 as the primary auditor and a small open-weight judge pilot (200 pairs); Gemini 3.1 Pro MT results are diagnostic rather than a full multi-LLM leaderboard. 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) 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. In the future, evaluation with paid annotators for each of the 20 languages could be a solution.

Use of AI Assistants

We used AI tools such as ChatGPT and Grammarly to correct grammatical errors, enhancing the coherence of the final manuscript and experimental scripts. While these tools have augmented our capabilities and contributed to our findings, it's pertinent to note that they have inherent limitations. We have made every effort to use AI in a transparent and responsible manner. Any conclusions drawn are a result of the first machine and then human combined insights.

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.

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*.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 4.

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. 2024a. 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.

Zhongtao Liu, Parker Riley, Daniel Deutsch, Alison Lui, Mengmeng Niu, Apurva Shah, and Markus Freitag. 2024b. Beyond Human-Only: Evaluating Human-Machine Collaboration for Collecting High-Quality Translation Data. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1095–1106, Miami, Florida, USA. 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.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The Portable Text Annotation Tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337, Abu Dhabi, UAE. Association for Computational Linguistics.

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

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

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 8.

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 C.

B Additional Details on Error Types

Table 8 shows the seven categories of machine translation error types studied in this work.

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

C Annotation Tool UI

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

D LLM-as-a-Judge Prompts

E Additional detail statistics on SSA-COMET-QE score

Figure 6 shows statistics for instances with scores before and after the quality qudit.

F Translation Error Labeling Agreement Between Human vs LLMs

Figure 7 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.

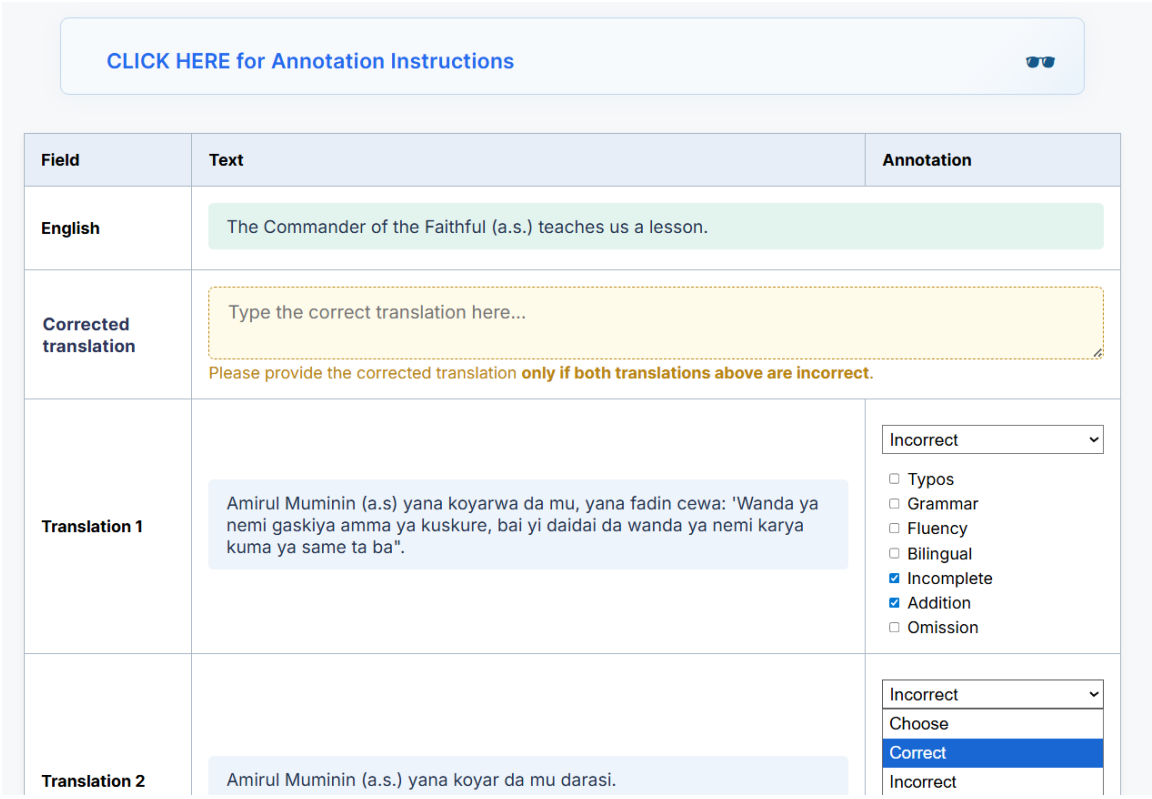

Figure 4: **Annotation tool UI for Hausa language**. The tool will be publicly released upon acceptance of the work for further machine translation and other NLP dataset quality audits with additional features.

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.
```

Figure 5: LLM-as-a-judge prompt details for translation quality analysis,

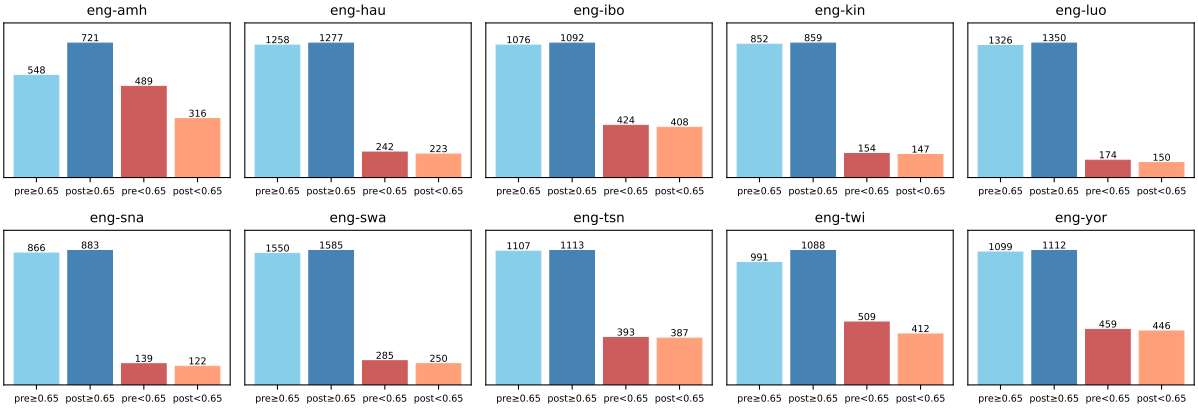

Figure 6: **SSA-COMET-QE translation scores across language pairs before (pre) and after (post) applying quality audit.** SSA-COMET-QE Scores range from 0 to 1, with higher values indicating better translation quality. The `en-pcm` direction has lower COMET scores; only 21% of the data have>0.5 SSA-COMET-QE because the language is not included in the specifically fine-tuned model. In all language cases, the number of instances increases from *pre* (SSA-COMET-QE score on the original data) to *post* (SSA-COMET-QE after human verification).

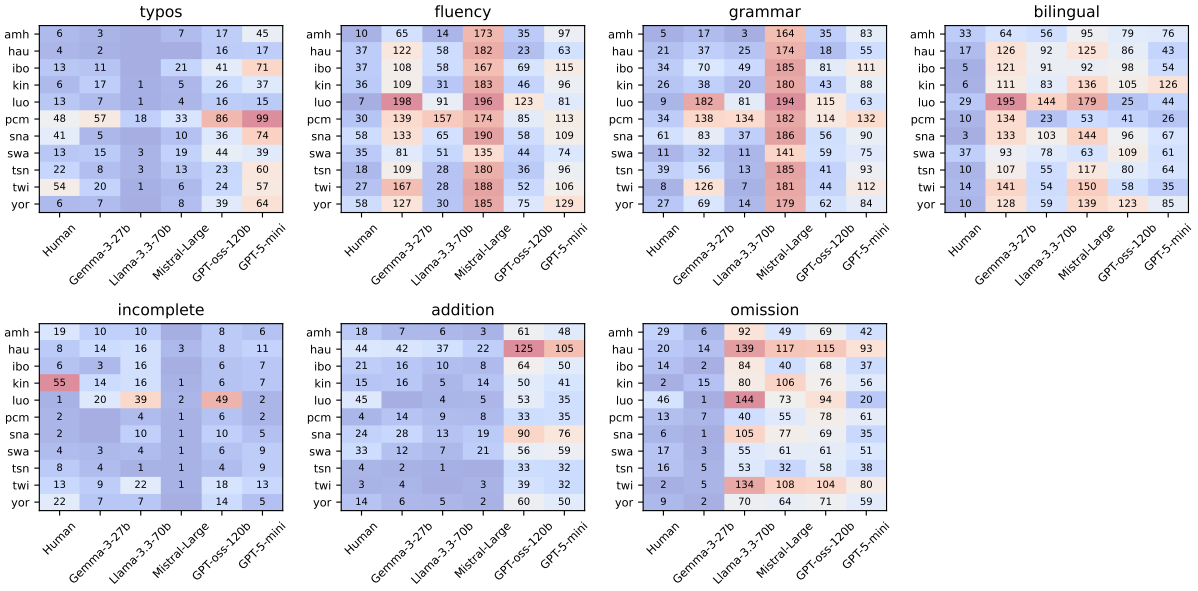

Figure 7: Translation error labeling overlap between Human and LLM-as-a-judge in 200 samples across languages.