AfriMMT-EA: Multi-domain Machine Translation for Low-Resource East African Languages

Naome A. Etori[1*†], **Kelechi Ezema**[2], **Nathaniel R. Robinson**[3], **Davis David**[4], **Alfred Malengo Kondoro**[5], **Elisha Ondieki Makori**[6], **Michael S. Mollel**[7*], **Maria L. Gini**[1†]
[1]University of Minnesota – Twin Cities, USA; [2]University of Colorado Boulder, USA; [3]Johns Hopkins University, USA; [4]Black Swan, Tanzania; [5]Hanyang University, South Korea; [6]University of Nairobi, Kenya; [7]Sartify Company Limited, Tanzania
{etori001,gini}@umn.edu

Abstract

Despite impressive performance and expanded language coverage in recent multilingual machine translation (MMT) systems, most African, and particularly East African, languages remain severely underrepresented in natural language processing (NLP) benchmarks, corpora, and state-of-the-art models (SOTA). With more than 2,000 languages spoken across Africa, current MT resources targets only a small fraction, leaving many languages underserved. To address this gap, we introduce **AfriMMT-EA**, the first large-scale multilingual MT dataset covering 53 East African languages across diverse domains. We expand low resource coverage by providing high-quality parallel data for many languages with no prior digital presence, including 23 new Kenyan and Tanzanian language pairs. We fine-tune **Gemma-3-270M** and **Gemma-3-1B**, to create our regionally adapted models **Safari-270M** and **Safari-1B** observing consistent translation quality improvements over strong off-the-shelf baselines, with the 1B model consistently outperforming the 270M variant across most languages. We release the dataset, trained models, and tools to lower barriers for researchers and communities, supporting more inclusive, robust, and culturally grounded MT research. All artifacts are publicly available at [1]

1 Introduction

The pursuit of eliminating global language barriers has positioned machine translation (MT) at the forefront of NLP research. Recent progress in multilingual MT has been driven by large language models such as GPT-4 (Hurst et al., 2024; Achiam et al., 2023; OpenAI, 2023), o1 (Jaech et al., 2024) LLaMA (Touvron et al., 2023; Dubey et al., 2024), Mistral (Chaplot, 2023), Qwen3 (Yang et al., 2025), Gemma (Team et al., 2025), BLOOM (Le Scao et al., 2023), and expanding multilingual corpora

Figure 1: Safari: Multilingual MT for 53 East African Languages fine-tuned on AfriMMT-EA [2]

for low-resource languages (LRLs) (Adelani et al., 2022b; Etori et al., 2025). These models demonstrate strong cross-lingual capabilities and are now widely adapted for translation tasks (Guo et al., 2024; Li et al., 2024). However, benefits from these advances remain unevenly distributed (Emezue and Dossou, 2022a; Orife et al., 2020a; Adebara et al., 2023). Despite rising interest in African MT, most existing systems still prioritize high-resource languages (Costa-Jussà et al., 2022), providing limited support for East African languages (Akera et al., 2022a; Martinus and Abbott, 2019). See Table 1.

Over the past decade, African MT research has progressed through new models (Elmadany et al., 2024a; Costa-Jussà et al., 2022; Üstün et al., 2024; Conneau, 2019; Adelani et al., 2021; Tonja et al., 2024a; Buzaaba et al., 2025), benchmarks (Adelani et al., 2022b; Goyal et al., 2022; Reid et al., 2021a), and evaluation frameworks such as AfriCOMET (Wang et al., 2024). To date, East Africa continues to have limited language coverage, and existing pan-African benchmarks reinforce this gap by relying on a subset of commonly used languages, leaving many others, such as Tanzania dialects, entirely overlooked.

Recent research have explored regional-focused

[1]Project repository: https://github.com/NEtori21/Multilingual-Machine-Translation-for-East-Africa

Findings of the Association for Computational Linguistics: EACL 2026, pages 3459–3492

MT Resource	EA/Total	EA Languages
MAFAND-MT (Adelani et al., 2022b)	5/17	amh, kin, lug, luo, swa
FLORES-200/NLLB (Costa-Jussà et al., 2022)	12/75	amh, dik, kam, kik, kin, lug, luo, gaz, run, som, swa, tir
SALT (Akera et al., 2022b)	6/6	swa, lug, lgg, ach, nyn, teo
CCAligned (El-Kishky et al., 2020)	7/106	am, om, so, sw, ti, lug, kin
MT560 (Gowda et al., 2021)	19/96	lug, ach, nyn, luo, run, swa, som, orm, kin, kam, kik , amh, nyn, dik, koo
FLORES-101 (Goyal et al., 2022)	9/101	amh, lug, kam, luo, orm, run, som, swh, tir
AFROMT (Reid et al., 2021a)	2/8	run, swa
InkubaLM (Tonja et al.)	1/5	swa
Toucan (Elmadany et al., 2024a)	15/45	aar, ach, amh, gez, kin, lgg, lug, nyn, orm, som, swa, swc, teo, tir, wal
Cheetah (Adebara et al., 2024)	10/517	ach, teo, lgg, lug, nyn, swa, som, tir, orm, run
mBART (Tang et al., 2020)	1/50	swh
mT0 (Muennighoff et al., 2022)	5/101	amh, orm, som, swh
AfriMMT-EA (Ours)	**53/53**	kln, dug, sxb, dav, guz, coh, teo, mas, cuh, pko, lsm, mer (See Table 3.)

Table 1: Comparison of East African coverage in existing MT datasets and models.

model approach as an alternative approach to global multilingual coverage. SEA-LION (Ng et al., 2025), SeaLLM (Nguyen et al., 2023), Sailor (Dou et al., 2024) demonstrate the effectiveness of such models for Southeast Asian (SEA) languages by extending multilingual open models LLaMA (Touvron et al., 2023) and Qwen (Bai et al., 2023). In Africa, InkubaLM (Tonja et al., 2024a) and Sunflower (Akera et al., 2025) extend this, showing the potential of localized modeling.

To address this gap, we adopt a region-focused localized modeling approach for East African languages, we introduce AfriMMT-EA, a large scale multilingual, multi domain MT benchmark covering 53 languages across Kenya, Uganda, Tanzania, Rwanda, and Ethiopia. We fine tune Gemma-3-(270M and 1B) (Team, 2025a) on all 714,490 translation pairs across 66 uni-directions (43M+ tokens). Our models achieve SOTA performance compared to the base model. We evaluate our models against the Open LLM Leaderboard [3] We summarize the contribution of our paper as follows:

- We introduce **AfriMMT-EA**, the first large-scale **region-focused**, **multi-domain** MT dataset for East Africa. Covering **53** local languages curated from diverse sources and partially translated by native speakers. Detailed breakdown of our contributions in terms of token counts, translation pairs, and domain coverage, in table 9

- We expand MT research language coverage by introducing **11 Kenyan and 12 Tanzanian** local languages into mainstream MT landscape, including **one endangered language, one extinct language**, and one language with no prior digital NLP support [4]

- We fine-tune **Gemma-3-270M** and **Gemma-3-1B**, release our models, dubbed as **Safari-270M** and **Safari-1B**, on **AfriMMT-EA** and providing a comprehensive evaluation across all 53 languages, where our regionally adapted safari models consistently outperform their Gemma base models and strong open multilingual off-the-shelf baselines.

2 Related Work

2.1 Multilingual MT for African Languages

MT research has historically prioritized high-resource languages (Reid et al., 2021a; Robinson et al., 2024; Workshop et al., 2022), leaving African languages underrepresented due to limited training and evaluation data (Adelani et al., 2022a; Etori and Gini, 2024; Ojo et al., 2023), with many still lacking reliable MT systems (Akera et al., 2022a). Despite broader coverage efforts (Nekoto et al., 2020; Bayes et al., 2024), substantial resource gaps persist. Existing tools such as Google Translate[5], Masakhane MT[6] (Orife et al., 2020b) and prior multilingual corpora (Agic and Vulic, 2019) offer only partial support, with Swahili remaining the most developed East African language for MT (De Pauw et al., 2009). Recent multilingual LLMs (Achiam et al., 2023; Team et al., 2023; Cui et al., 2025) still underserve many East African languages (e.g., Bena, Hehe, Suba). This gap has motivated African-centered benchmarks and models such as IrokoBench (Adelani et al., 2024), AfroBench (Ojo et al., 2023), AfriMTE and AfriCOMET (Wang et al., 2023), InkubaLM (Tonja et al., 2024a), SALT (Akera et al., 2022b), Sunflower (Akera et al., 2025), Lugha-Llama (Buzaaba et al., 2025), Toucan (Elmadany et al., 2024a; Reid et al., 2021b), AfroLingu-MT (Elmadany et al., 2024b), AfroXLM-R (Alabi et al., 2022),

[3] https://github.com/hiyouga/LlamaFactory.

[4] Language classification and vitality information were obtained from https://www.ethnologue.com/.

[5] https://translate.google.com

[6] https://www.masakhane.io/

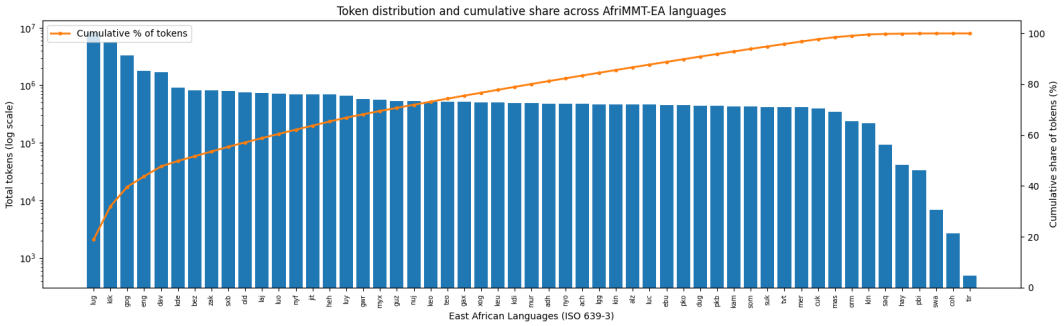

Figure 2: Token distribution across **AfriMMT-EA**. Blue bars show total token counts per language on a log scale, using ISO codes. The orange line indicates the cumulative share of tokens across languages.

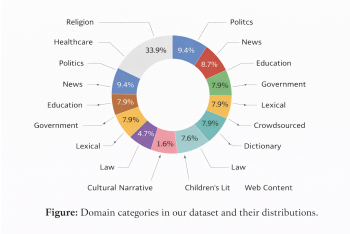

Figure: Domain categories in our dataset and their distributions.

Figure 3: Domain categories in our dataset and their distributions.

and AfroXLMR-Social. Regionally focused trends, including Sunflower (Akera et al., 2025), SEA-LION (Ng et al., 2025), and SeaLLM (Nguyen et al., 2023), further emphasize localized modeling. Our work advances this direction by explicitly targeting East African languages such as Tanzanian dialects, which remain absent in prior MT research.

3 Methodology and Dataset

Our curation pipeline emphasizes data quality (Oladipo et al., 2023) through careful examination and selection of existing sources.

3.1 Language selection criteria

We selected languages from five East African countries, as outlined in §1, guided by four motivations. **First**, our team includes **five** native speakers from the region, specifically **three** from Tanzania and **two** from Kenya, whose linguistic expertise informed the selection. **Second**, as noted in §2, LRL data collection, annotation and alignment are resource-intensive, and training data-hungry models is costly. We therefore prioritized languages with available datasets, with collectors using their best judgment in sourcing materials. **Third**, current LLMs and benchmarks include only a few African languages such as Swahili, Hausa, Yoruba, and Amharic (Akera et al., 2025). AfriMMT-EA instead follows a region-first, language-centric strategy that centers East African languages as core research focus. **Fourth**, the five countries share **English**, **Swahili**, and many indigenous languages, which aligns with our uni-direction translation setup. Because our benchmark includes limited **English→local** data but far more **Swahili→local** pairs, Swahili serves as a strong pivot for cross-lingual transfer

3.2 Language characteristics

Most languages in AfriMMT-EA use Latin-based orthographies, ranging from standard Latin script such as Kikuyu, Suba, Luganda to Latin Extended and modified Latin scripts with long vowels, pre-nasalized consonants, diacritics, and glottal marking such as Tharaka, Marakwet, Borana, Samburu, Maasai. Tigrinya is written in the Ethiopic **Ge'ez** script and our dataset preserve that.

3.3 Data sources and composition

Regional Data Sources: We curated from multiple sources, initially we classified our data into five primary categories: (i) existing digital corpora, (ii) digitized physical or printed materials, (iii) specialized linguistic resources such as dictionaries, cultural narratives (dictionaries, glossaries, and cultural narratives); (iv) community and locally developed documents, such as native-speaker translations of existing texts; and (v) directly curated datasets, collected and annotated through our data curator platform shown in Figure 6.

3.4 Domain Composition and Distribution

Our corpus spans multiple domains, as summarized in Figure 3,Table 2 and Table 7. **Kenya** contributes the widest domain coverage, including politics, religion, short stories, cultural narratives, agriculture, folklore, African literature, lex-

ical resources, education, and healthcare. Asubstantial portion is drawn from *religious corpora*, primarily sourced from *Bible Translation & Literacy (BTL)* and *Bible.com*[7]. The data are complemented by digitized books, PDFs, and cultural archives processed through a Gemini-GUI, and by publicly available publications and research corpora (Mbogho et al., 2025; Wanjawa et al., 2022a; Adelani et al., 2022b; Team et al., 2022; Tiedemann, 2012). **Tanzania.** As shown in Table 2, Tanzanian data constitute 14.4% of the corpus and are drawn primarily from religious[7], lexical, and crowdsourced. The crowdsourced portion was collected through a controlled web-based translation interface, where native speakers submitted and validated translations As shown in Figure 6.

3.5 Data collection

We divided our data collection methodology into three main steps: *searching and gathering*, *extraction* and *alignment*. We then followed a structured data collection procedure. **(i)** First we found existing resources by querying research databases and search engines. Specifically,we searched for the language name in the *ACL Anthology*[8] and used combinations of keywords such as "`[language name]` machine translation", "`[language name]` NLP", "`[language name]` Translation", and "`[language name]` African MT" on Google Scholar[9]. We widen search for parallel corpora using the Lanfrica [9] database and general Google search engine [9]. Given the large number of query results, we focused on the most relevant entries. **(ii)** Second, we searched online for books , PDFs offering translations or supplementary language resources, including the Internet Archive.[9] **(iii)** Third, we contacted researchers and community members from the languages' communities to identify potential data sources such as BTL [7] While response rates were low, these contacts enabled us to obtain some otherwise unavailable datasets. **(iv)** Fourth, we engaged native speakers by posting calls for contributions to translation tasks on social media such as Facebook and LinkedIn. This yielded some additional data, though many participants were unable to complete the translations fully. **(v)** Fifth, we established a web-based platform, *Argilla*, that enables East Africans native speakers to voluntarily contribute to MT data for their own languages[17]. As shown in Figure 6 and Figure 8. Complete URLs for all external resources mentioned in this section are listed in Appendix J.

Data Extraction: After data gathering, we categorized the resources into seven types: **(i)** previously published parallel datasets, **(ii)** web sources with pre-aligned parallel sentences, **(iii)** web sources containing full-text articles accompanied by translations, **(iv)** PDFs with aligned parallel sentences, **(v)** human-translated data generated via the Argilla platform or native speaker contributors, **(vi)** Bible parallel translations, and **(vii)** other PDF sources. Previously published parallel datasets were combined at the beginning of the extraction process, including single-language resources such as KenTrans (Wanzare et al., 2022), Kencorpus (Wanjawa et al., 2022b), Lumasaba (Nabende et al., 2023), (Mbogho et al., 2025) and (Wanjawa et al., 2022a) . We customized extraction procedures for PDFs, websites, and other sources, leveraging *Gemini 2.5 Flash* (Comanici et al., 2025) [10], with prompt-driven workflows followed by human-in-the-loop quality assurance (See Fig. 4.). The extraction and annotations team included two Kenyan native speakers, three Tanzanian native speakers, and one Nigerian collaborator with linguistic familiarity in Ugandan languages. Although this did not cover all included languages, the annotators' knowledge of related regional languages ensured accuracy, supported by careful quality control.

Data Alignment: The alignment process was crucial for transforming raw bilingual data into structured parallel data suitable for our downstream MT task. For web and document-based sources containing unsegmented text and their translations, we employed LLM-assisted segmentation and alignment with human verification. First, we copied text from the target languages directly from the original source text pair into the *Gemini 2.5 Flash* (Comanici et al., 2025) user interface, prompting it to identify and align corresponding sentence boundaries across source and target languages. The model segmented text based on punctuation, paragraph boundaries, and translation cues such as repetition or discourse markers. Next came our human validation phase. Annotators manually inspected the aligned segments to correct mismatches, missing sentences, or partial alignments. After verification, all aligned sentence

[7]Source URLs: `btlkenya.org`, `bible.com`, `hf.co/datasets/allandclive/UgandaLex2`

[8]Source URLs: `https://aclanthology.org`, `https://scholar.google.com`, `https://lanfrica.com/`

[9]`https://archive.org/`

pairs were organized and cleaned using Excel or Google Sheets and stored in comma-separated values (CSV) format.

3.6 Data preprocessing pipeline

Our Preprocessing pipeline was developed to construct a high-quality parallel multilingual translation dataset. The pipeline follows a systematic sequence of steps: **(i)** data ingestion from heterogeneous CSV sources with automatic encoding detection, **(ii)** comprehensive text cleaning to remove encoding artifacts, normalize whitespace and punctuation, and filter empty or invalid entries, **(iii)** de-duplication of exact translation pairs, **(iv)** tokenization and metadata enrichment, **(v)** concurrent processing for efficient handling of multiple languages, **(vi)** deterministic train-test splitting per language pair, **(vii)** quality assurance and statistical validation, and **(viii)** dataset export and publication, including a comprehensive README with statistics and language coverage for reproducibility.

Data Organization and Ingestion: The raw dataset was organized hierarchically by country and language, with each language pair stored in separate CSV files. We employed automatic encoding detection to handle files with varying character encodings *UTF-8, UTF-8-BOM, Latin-1, CP1252, ISO-8859-1*, ensuring proper text extraction regardless of original file format. This multi-encoding support was critical given the diverse sources and historical nature of some texts. Files with missing orcontent were logged for inspection.

Text Cleaning: To address common data quality issues in our corpora, each sentence was processed in a multi-stage cleaning procedure. The cleaning process began with systematic removal of encoding artifacts/erroneous UTF-8 symbols (e.g. $\surd$, ¬, ó)), byte-order marks, non-breaking spaces, and zero-width character artifacts from file format conversions and digitization processes. Regular expressions were used to identify and remove patterns such as unique accented characters, negation artifacts, and malformed Unicode sequences. Next, we normalized whitespace and punctuation, removed standalone accents, converted ASCII, and removed translation pairs shorter than 3 characters or consisting solely of punctuation marks. We counted tokens using the Gemma-3 tokenizer to track source and translation token counts for each pair, enabling analysis of translation length ratios and informing model training decisions.

Country	Languages	Translation Pairs	Percentage
Kenya	24	351,396	48.8%
Uganda	16	256,133	35.5%
Tanzania	12	103,520	14.4%
Rwanda	1	6,246	0.9%
Ethiopia	2	3,441	0.5%

Table 2: **Geographic Distribution** of translation pairs by country. Kenya and Uganda dominate the dataset, together accounting for over 80% of the total data.

Deduplication and Dataset Splitting: To prevent data leakage and ensure model generalization, we implemented rigorous de-duplication. We removed translation pairs with identical source and translation text, and logged duplicate entries for inspection. The de-duplication process was applied globally across all languages to eliminate cross-contamination between language pairs. For dataset splitting, we employed a stratified approach that maintains the distribution of language pairs in both training and test sets. Each language pair was independently shuffled using a fixed random seed (42) for reproducibility, then split using an 80/20 ratio. A similar approach that maximizes data coverages was employed for languages with fewer pairs. This stratified splitting ensures every language pair is adequately represented in both splits while maintaining evaluation integrity. The final dataset structure has **source language**, **target language**, **country of origin**, **token counts**, **processing timestamp**, and **processor version** for traceability. Statistical validation confirmed zero remaining duplicates post-processing and verified the 80/20 split ratio across all language pairs.

4 Experimental Setup

We leverage pretrained instruction-tuned language models as our starting point, hypothesising that their inherent instruction-following capabilities will facilitate effective transfer learning for machine translation tasks. Our model selection strategy prioritises inference efficiency suitable for edge deployment and multilingual capabilities inherent in the tokenizer. Consequently, we focused on lightweight models, primarily targeting the sub-1B parameter range while including select compact models for comparison. The candidate models include the Gemma (270M and 1B) (Team, 2025a), Llama 3.2 (1B) (Grattafiori et al., 2024), Qwen 2.5 (0.5B) (Team, 2024; Yang et al., 2024), Phi-4-mini (3.8B) (Xu et al., 2025), and SmolLM2-Instruct

ISO Code	**Name**	Family	Region	# Speakers	Vitality	Digital support
swa	Swahili	Niger_Congo/Bantu	East and Central Africa	87.2 M	Institutional	Vital
dug	Duruma	Niger-Congo/Bantu	Kenya	600K	Institutional	Ascending
sxb	Suba	Niger-Congo/Bantu	Kenya and Tanzania	140K	Stable	Ascending
dav	Taita	Niger-Congo/Bantu	Kenya	370K	Stable	Ascending
mer	Meru	Niger-Congo/Bantu	Kenya	2M	Institutional	Ascending
guz	Gusii	Niger-Congo/Bantu	Kenya and Tanzania	2.7M	Stable	Ascending
lsm	Saamia	Niger-Congo/Bantu	Kenya and Uganda	480K	Institutional	Emerging
kam	Kamba	Niger-Congo/Bantu	Kenya	5.2M	Institutional	Ascending
teo	Teso	Nilo-Saharan/Nilotic	Kenya and Uganda .	2.78M	Institutional	Ascending
kln	Kalenjin	Nilo-Saharan/Nilotic	Kenya	6.6M	Stable	Ascending
ebu	Embu	Niger-Congo/Bantu	Kenya	320K	Stable	Ascending
coh	Chonyi	Niger-Congo/Bantu	Coastal Kenya	310K	Stable	Still
lwg	Luwanga	Niger-Congo/Bantu	Western Kenya	103K	Stable	Emerging
kik	Kikuyu	Niger-Congo/Bantu	Central Kenya	8.3M	Stable	Ascending
luo	Luo	Nilo-Saharan/Nilotic	Kenya and Tanzania	4.2M	Stable	Ascending
mas	Maasai	Nilo-Saharan/Nilotic	Kenya and Tanzania	1.5M	Stable	Ascending
gax	Borana	Afro-Asiatic/Cushitic	Kenya, Somalia and Ethiopia	9.6M	Stable	Ascending
thk	Tharaka	Niger-Congo/Bantu	Kenya	300K	Institutional	Ascending
bxk	Bukusu	Niger-Congo/Bantu	Kenya and Uganda	1.4M	Institutional	Ascending
rag	Maragoli	Niger–Congo/Bantu	Kenya	620K	Institutional	Ascending
cuh	Chuka	Niger–Congo/Bantu	Kenya	250K	Stable	Emerging
saq	Samburu	Nilo-Saharan	Kenya	240K	Stable	Ascending
enb	Marakwet	Nilo-Saharan	Kenya	180K	Stable	Ascending
tvs	Taveta	Niger–Congo/Bantu	Kenya	21K	Stable	Emerging
nyf	Giriama	Niger–Congo/Bantu	Kenya	1M	Institutional	Ascending
pkb	Pokomo	Niger–Congo/Bantu	Kenya	95K	Institutional	Ascending
pko	Pokot	Nilo-Saharan	Kenya	700K	Stable	Ascending
som	Somali	Afro-Asiatic	Kenya and Somalia	24M	Institutional	Vital
gax	Oromo	Afro-Asiatic	Ethiopia	45.5M	Stable	Ascending
tir	Tigrinya	Afro-Asiatic	Ethiopia	9.9M	Institutional	Vital
kin	Kinyarwanda	Niger-Congo/Bantu	Rwanda	15M	Institutional	Vital
bez	Bena	Niger-Congo	Central Tanzania	590K	Stable	Ascending
gog	Gogo	Niger-Congo	Central Tanzania	1.4M	Stable	Ascending
kde	Makonde	Niger-Congo	SE. Tanzania, No. Mozambique	2.1M	Stable	Ascending
lag	Langi	Niger-Congo	Tanzania	410K	Stable	Ascending
mas	Maasai	Nilo-Saharan	Tanzania, Kenya	2.1M	Stable	Ascending
jit	Jita	Niger-Congo	Tanzania	210K	Stable	Ascending
heh	Kihehe	Niger-Congo	Southern Tanzania	1.2M	Endangered	Ascending
old	Kichaga	Niger-Congo	Kilimanjaro, Tanzania	1.4M	Stable	Emerging
suk	Kisukuma	Niger-Congo	Northern Tanzania	8.1M	Stable	Ascending
hay	Kihaya	Niger-Congo/Bantu	Northwestern Tanzania	1.3M	Stable	Ascending
asu	Pare	Niger-Congo	Northeast Coast Tanzania	500K	Stable	Ascending
ach	Acholi	Nilo-Saharan	Uganda, South Sudan	2M	Institutional	Ascending
alz	Alur	Nilo-Saharan	DRC, Uganda	2.5M	Institutional	Ascending
luc	Aringa	Nilo-Saharan	Uganda	927K.	Stable	Ascending
teo	Ateso	Nilo-Saharan	Uganda, Kenya	3.1M	Institutional	Ascending
gwr	Gwere	Niger-Congo	Uganda	876K	Institutional	Ascending
adh	Jopadhola	Nilo-Saharan	Uganda	570K	Stable	Ascending
keo	Kakwa	Nilo-Saharan	Uganda, DRC, South Sudan	591K	Stable	Ascending
ndp	Kebu	Nilo-Saharan	Uganda	51.7K	Stable	Emerging
kdi	Kumam	Nilo-Saharan	Uganda	350K	Institutional	Emerging
laj	Lango	Nilo-Saharan	Uganda	2.7M	Institutional	Ascending
lgg	Lugbara	Nilo-Saharan	Uganda, DR Congo	1.2M	Institutional	Ascending
myx	Masaaba	Niger-Congo	Uganda	2.1M	Institutional	Ascending
nuj	Nyole	Niger-Congo	Uganda	633K	Stable	Ascending
nyo	Nyoro	Niger-Congo	Uganda	1.2M	Institutional	Ascending
xog	Soga	Niger-Congo	Uganda	3.7M	Institutional	Ascending

Table 3: ISO Language Codes, Names, Language Families, Regions (where the languages are spoken in Africa), Speaker Estimates, Vitality, and Digital Language Support, per Ethnologue (https://www.ethnologue.com).

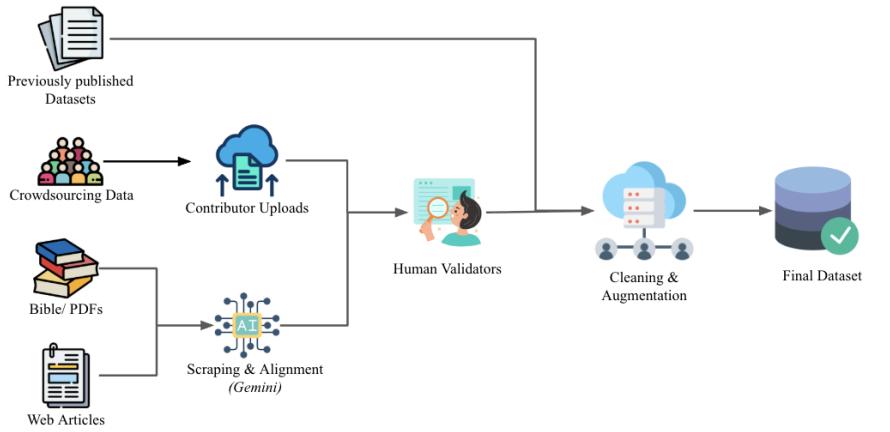

Figure 4: Data Collection Pipeline. Our framework integrates resource search, text extraction, and dual alignment (crowdsourced + Gemini-assisted) to produce validated East African multilingual data.

tokenizer	Languages	fertility	PCW
Gemma-3	All	1.9683	58.44%
Llama-3.2	All	2.1153	60.73%
Phi-4	all	2.1330	61.16%
Qwen-3	all	2.1494	62.25%
SmolLM2-1.7B-Instruct	all	2.2596	63.98%

Table 4: Comparison of tokenizer efficiency metrics (Fertility and PCW) across candidate base models, aggregated across all target languages.

(1.7B) (Allal et al., 2025). Given that our methodology relies strictly on Supervised Fine-Tuning (SFT) rather than pre-training, we assessed the models' tokenization efficiency using two metrics: Fertility (average tokens per word) and Proportion of Continued Words (fragmentation rate) (Penedo et al., 2025). As shown in Table 4, these metrics guided our final selection. Due to computational constraints preventing a full training/evaluation of all candidates, we proceeded with the Gemma family as it offered the optimal balance of tokenizer efficiency and model size.

4.1 Task Overview

We formulate AfriMMT–EA as a supervised MT task, where the model learns from parallel sentence pairs (x, y) drawn from the Swahili→local and English→local translation directions. Each training example is converted into a structured three-turn conversation following the Gemma-3 chat template: (1) a system instruction describing the translation task, (2) a user message specifying the translation direction and source text, and (3) an assistant response containing the target translation, see prompting strategy at Appendix I. Because every instance includes explicit `source_language` and `translated_language` fields, the model never needs to infer the target language. During fine-tuning, we use response-only training, applying loss exclusively to the assistant turn and masking all system and user tokens. At inference time, the model receives the same two-turn (system + user) prompt and must produce the target translation without access to the reference.

4.2 Baseline and Evaluation Models

We employ Gemma-3-(270M and 1B) (Team, 2025a) as baseline models, chosen for their strong multilingual,vocabulary capacity and lightweight design, which allows full fine-tuning and deployment in low-resource environments. To contextualize the performance of our fine-tuned models, we evaluate against open LLMs, including **Meta-Llama-3.3 1B** (Dubey and et al., 2024), **Qwen3** family (0.6B and 1.7B) (Team, 2025b). All models are decoder-only that have demonstrated strong multilingual performance.

4.3 Evaluation Metrics

Evaluation uses the test split from the same dataset, filtered to match the trained language direction. We compute three complementary metrics: BLEU (Papineni et al., 2001), chrF++ (Popović, 2017), and TER (Snover et al., 2006), following common MT practice. All metrics are calculated using the sacrebleu library (Post et al., 2025) with international tokenization. Results are aggregated at the level of individual language pairs, and weighted by the number of test examples to compute overall averages. We record precision breakdowns, brevity penalty,

and average system versus reference lengths.

4.4 Prompting Strategy

As explained in §4.1. We use a fixed ChatML template consisting of: (i) a system instruction defining the task, (ii) a user message containing the source, and target language, input sentence, (iii) the assistant translation.(See §I).

4.5 Hyperparameter optimization

The model was configured with a maximum sequence length of 2,048 tokens to accommodate longer translations. Hyperparameter optimization was conducted using Weights & Biases sweeps to identify optimal training configurations. The training process used the AdamW 8-bit optimizer for memory efficiency, with batch sizes and gradient accumulation steps varied across experiments. We employed a linear learning rate schedule with 5 warmup steps. The learning rate, weight decay, number of training epochs, and effective batch size (product of per-device batch size and gradient accumulation steps) were treated as hyperparameters subject to optimization. All experiments used gradient checkpointing via the Unsloth framework (Han et al., 2023) to reduce memory consumption during training. Translations were generated using greedy decoding (temperature=0.0) with a maximum length of 128 tokens. We employed batch inference with a batch size of 256 to accelerate evaluation, processing the entire test set efficiently.

4.6 Supervised fine-tuning

We perform supervised fine-tuning (SFT) on two base models, Gemma-3-270M and Gemma-3-1B (Team, 2025a) described in §4.2, to specialize them for AfriMMT-EA language in 66 language directions. Each model is trained using a chat-formatted parallel corpus with sentence alignment.

4.7 Inference Setting

At evaluation time, the model is switched to `eval` mode and queried using the same chat-style prompting schema as in training. For each test instance, we construct a prompt that specifies the translation direction along with the source text. Translations are generated using greedy decoding (temperature = 0.0) with a maximum of 128 new tokens. To improve throughput, we perform batched inference: prompts are grouped into mini-batches of size 16, tokenized, and decoded on device. For each output, the generated tokens following the prompt span are extracted and detokenized to obtain the hypothesized translation.

4.8 Logging and Artifact Reproducibility

All scores are saved to disk and logged to Weights & Biases [10]. Aggregated predictions and metrics per language pair are stored as W&B artifacts for reproducibility.

5 Results and discussion

5.1 Impact of Regional Fine-Tuning

Figure 5 summarizes chrF++ and BLEU performance across 66 English→local and Swahili→local translation directions, comparing the base Gemma models with their regionally fine-tuned Safari variants and other open-source baselines. Regional fine-tuning achieves large improvements over the base Gemma models across nearly all language pairs. Both Safari-270M and Safari-1B substantially outperform their base counterparts, with Safari-1B achieving the best performance and consistent improvement overall. In particular, for Swahili→local translation, Safari-1B achieves chrF++ scores above 35 for the majority of language pairs, and exceeds 50 chrF++ for several high-performing directions such as Swahili→Luganda, Swahili→Giriama, Swahili→English, and Swahili→Kakwa. In contrast, the Gemma base models achieve chrF++ scores below 20 for over 80% of Swahili→local directions, with a substantial fraction remaining in the single-digit range.

5.2 Language Directional Asymmetry

Figure 5 shows a clear and consistent directional asymmetry between English→local and Swahili→local translation, with Swahili→local translation substantially outperforming English→local translation across languages. This pattern should be interpreted in light of the training data distribution: the models are exposed to substantially more Swahili→local parallel data than English→local, biasing learning toward Swahili as a more effective source pivot. Even under this constraint, the magnitude of the asymmetry is pronounced. For example, Swahili→English achieves 55.26 chrF++ / 33.14 BLEU (Safari-1B), whereas English→Swahili remains far weaker at 15.28 / 0.50, indicating that the model is significantly better at generating English given

[10] https://wandb.ai/site/

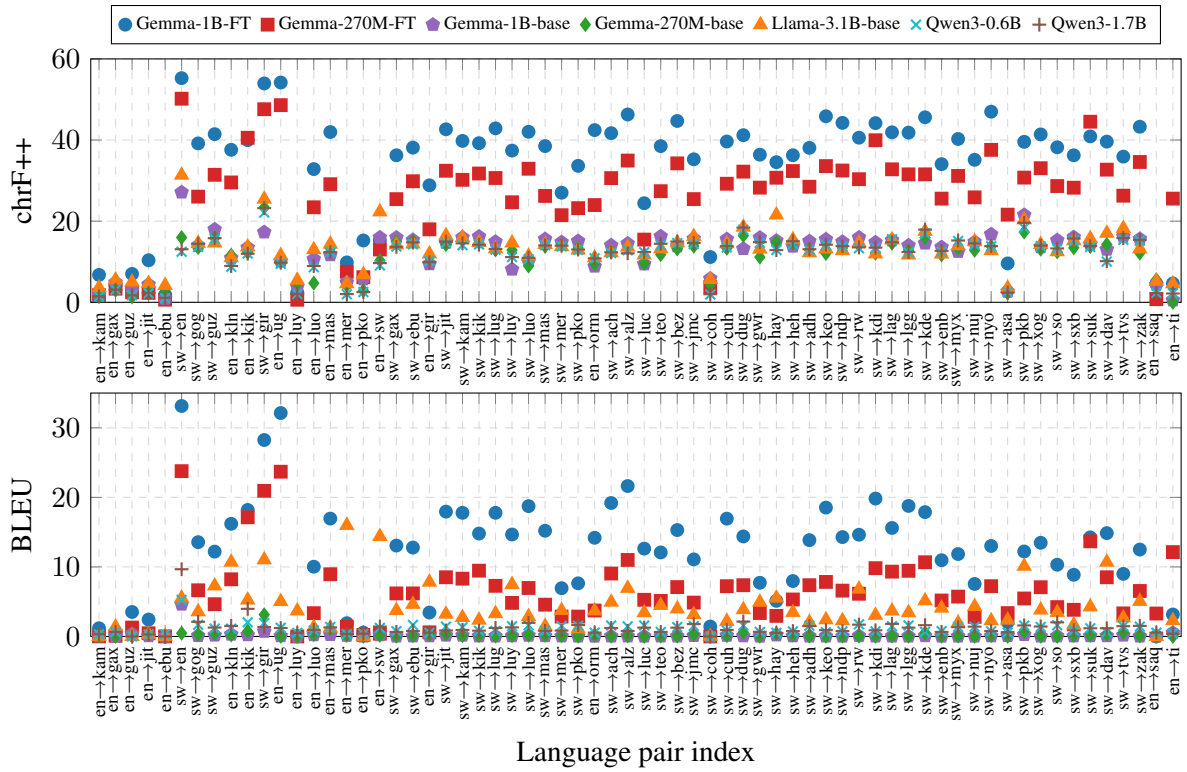

Figure 5: chrF++ and BLEU models scores for English→local and Swahili→local across 66 language direction.

Swahili than the reverse. Luhya shows near-zero performance in the English direction **(En→Luy: 2.43 chrF++, 0.03 BLEU)** but performs better from Swahili **(Sw→Luy: 37.40 / 14.66)**. Similar directional disparities appear for **Giriama (25.84 / 3.44 vs. 53.95 / 28.25)** and **Luo (32.84 / 10.03 vs. 42.04 / 18.94)**. Among En→local directions, languages with better training coverage (e.g., En→Lug, En→Kik) achieve comparatively higher scores.

5.3 Model size comparison

Across the 66 translation directions in Figure 10 and Figure 11 Safari-1B model has the best performance, followed by Gemma-3-270M-FT, both of which exceed the Qwen baselines by wide margins in BLEU and chrF++. Specifically, **Safari-1B** attains a mean chrF++ of **33.65** compared to **26.17** for **Safari-270M**, and the same ordering holds for BLEU (mean **12.06** vs. **6.44**). Llama-3-1B consistently outperforms both Qwen3-0.6B and Qwen3-1.7B, with the largest gaps appearing on "high-resource". For instance, in **En↔Sw**, Llama-3-1B yields substantially higher BLEU and chrF++ than either Qwen model, and the same pattern holds for other higher-test-sample directions such as **Kikuyu↔English** and **Swahili→Giriama**. Qwen3-1.7B reliably improves over Qwen3-0.6B, but these gains remain modest and do not close the large quality gap to Llama-3-1B.

6 Conclusion

Despite Africa's large population and rich linguistic diversity, multilingual models continue to offer limited coverage and uneven support for many African languages. In this work, we introduced **AfriMMT-EA**, a large-scale multi-domain MT dataset covering 53 East African languages across **66 translation directions**. Using AfriMMT-EA, we fine-tuned GEMMA 3-270M and 3-1B models to produce East African regionally focused models, which we release as **Safari-270M** and **Safari-1B**, establishing strong MT baselines for East African languages. Our results demonstrate that region-specific data curation and fine-tuning substantially improve translation quality over base multilingual models, particularly for **Swahili→local** directions and common regional dialects. Our Safari models and datasets are publicly released to support Machine Translation Research for Low Resource languages. We underscore the need to expand LRL data for region-focused datasets, both as evaluation benchmarks and as finetuning resources for culturally inclusive MT systems.

Limitations and Future Work

1. Our study covers 53 languages from five East African countries (Kenya, Uganda, Tanzania, Ethiopia, and Rwanda). Expansion to additional countries was constrained by data availability and curation resources, positioning AfriMMT-EA as a foundational step for future regional MT extensions.

2. Model training was limited by computational and financial resources. As a result, we fine-tuned only two Gemma-3 models and evaluated open-source systems, meaning observed performance gaps may not capture the full multilingual MT landscape. Broader model comparisons and ablations remain key future directions.

3. Although we rely on publicly released datasets under CC BY licenses with proper attribution, AfriMMT-EA includes more Swahili→local than English→local data, leading to directional performance asymmetries. Future work will expand English→local coverage to reduce this imbalance.

4. We acknowledge that BLEU and chrF may be unreliable for morphologically rich, LR East African languages, and thus may underrepresent true translation quality. Future work should incorporate human evaluations and semantic metrics to provide a more accurate assessment.

Ethics Consideration

East African languages have long been marginalized in linguistic research and digital technologies. Our work aims to promote equitable representation in multilingual NLP while acknowledging the complex sociocultural histories that shape language use in the region. We recognize that some communities may hold differing views on the value of MT and language modeling. While some may welcome these tools for education, digital access, and language preservation, others may be concerned about data misuse, cultural commodification, or dilution of linguistic identity. We believe the potential social benefits outweigh these risks, particularly in supporting LRLs development.

Our dataset includes cultural, religious, and historical texts that may reflect outdated or sensitive viewpoints. We do not endorse these perspectives but preserve them to maintain linguistic authenticity. All data were either publicly accessible or shared with consent. Although our collection represents many languages, it may not capture full dialectal or community variation, potentially introducing bias.

We encourage continued evaluation across dialects and domains and emphasize that these resources should not be used for harmful purposes. Finally, AI tools such as ChatGPT, Gemini were used to assist documentation; all outputs were reviewed to ensure accuracy, cultural respect, and compliance with the ACL Code of Ethics.

Acknowledgments

We would like to thank Sartify Company Limited and PAWA-AI for providing logistical support for data collection and access to computational resources. We also thank all volunteers, data contributors, annotators, and publishers for their support and for granting access to the materials used in this research.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Ife Adebara, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2024. Cheetah: Natural language generation for 517 african languages. *arXiv preprint arXiv:2401.01053*.

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2023. SERENGETI: Massively Multilingual Language Models for Africa. ArXiv:2212.10785 [cs].

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

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 Chinenye Emezue, Colin Leong, Michael Beukman, Shamsuddeen Hassan Muhammad, Guyo Dub Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ayoade Ajibade, Tunde Oluwaseyi Ajayi,

Yvonne Wambui Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Koffi Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022a. A Few Thousand Translations Go a Long Way! Leveraging Pre-trained Models for African News Translation. ArXiv:2205.02022 [cs].

David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, et al. 2022b. A few thousand translations go a long way! leveraging pre-trained models for african news translation. arXiv preprint arXiv:2205.02022.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba O Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, et al. 2024. Irokobench: A new benchmark for african languages in the age of large language models. *arXiv preprint arXiv:2406.03368*.

Zeljko Agic and Ivan Vulic. 2019. Jw300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. Available at: https://scholar.google.com/scholar?cluster=8239399398272024687&hl=en&as_sdt=0,24.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, et al. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Benjamin Akera, Jonathan Mukiibi, Lydia Sanyu Naggayi, Claire Babirye, Isaac Owomugisha, Solomon Nsumba, Joyce Nakatumba-Nabende, Engineer Bainomugisha, Ernest Mwebaze, and John Quinn. 2022a. Machine Translation For African Languages: Community Creation Of Datasets And Models In Uganda.

Benjamin Akera, Jonathan Mukiibi, Lydia Sanyu Naggayi, Claire Babirye, Isaac Owomugisha, Solomon Nsumba, Joyce Nakatumba-Nabende, Engineer Bainomugisha, Ernest Mwebaze, and John Quinn. 2022b. Machine translation for african languages: Community creation of datasets and models in uganda. In *3rd Workshop on African Natural Language Processing*.

Benjamin Akera, Evelyn Nafula Ouma, Gilbert Yiga, Patrick Walukagga, Phionah Natukunda, Trevor Saaka, Solomon Nsumba, Lilian Teddy Nabukeera, Joel Muhanguzi, Imran Sekalala, et al. 2025. Sunflower: A new approach to expanding coverage of african languages in large language models. *arXiv preprint arXiv:2510.07203*.

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

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. 2025. Smollm2: When smol goes big – data-centric training of a small language model.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Edward Bayes, Israel Abebe Azime, Jesujoba O Alabi, Jonas Kgomo, Tyna Eloundou, Elizabeth Proehl, Kai Chen, Imaan Khadir, Naome A Etori, Shamsuddeen Hassan Muhammad, et al. 2024. Uhura: A benchmark for evaluating scientific question answering and truthfulness in low-resource african languages. *arXiv preprint arXiv:2412.00948*.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, John D Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Happy Buzaaba, Alexander Wettig, David Ifeoluwa Adelani, and Christiane Fellbaum. 2025. Lugha-llama: Adapting large language models for african languages. *arXiv preprint arXiv:2504.06536*.

Devendra Singh Chaplot. 2023. Albert q. jiang, alexandre sablayrolles, arthur mensch, chris bamford, devendra singh chaplot, diego de las casas, florian bressand, gianna lengyel, guillaume lample, lucile saulnier, lélio renard lavaud, marie-anne lachaux, pierre stock, teven le scao, thibaut lavril, thomas wang, timothée lacroix, william el sayed. *arXiv preprint arXiv:2310.06825*, 3.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard,

et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study. *arXiv preprint arXiv:2502.02481*.

Guy De Pauw, Peter Waiganjo Wagacha, and Gilles-Maurice de Schryver. 2009. The sawa corpus: a parallel corpus english-swahili. In *Proceedings of the First Workshop on Language Technologies for African Languages*, pages 9–16.

Longxu Dou, Qian Liu, Guangtao Zeng, Jia Guo, Jiahui Zhou, Wei Lu, and Min Lin. 2024. Sailor: Open language models for south-east asia. *arXiv preprint arXiv:2404.03608*.

Abhimanyu Dubey and et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.12345*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kevin Duh, Paul McNamee, Matt Post, and Brian Thompson. 2020. Benchmarking neural and statistical machine translation on low-resource african languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2667–2675.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. Ccaligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969.

AbdelRahim Elmadany, Ife Adebara, and Muhammad Abdul-Mageed. 2024a. Toucan: Many-to-many translation for 150 african language pairs. *arXiv preprint arXiv:2407.04796*.

AbdelRahim Elmadany, Ife Adebara, and Muhammad Abdul-Mageed. 2024b. Toucan: Many-to-Many Translation for 150 African Language Pairs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13189–13206, Bangkok, Thailand. Association for Computational Linguistics.

Chris C. Emezue and Bonaventure F. P. Dossou. 2022a. MMTAfrica: Multilingual Machine Translation for African Languages. ArXiv:2204.04306 [cs].

Chris C Emezue and Bonaventure FP Dossou. 2022b. Mmtafrica: Multilingual machine translation for african languages. *arXiv preprint arXiv:2204.04306*.

Naome Etori and Maria Gini. 2024. Rideke: Leveraging low-resource twitter user-generated content for sentiment and emotion detection on code-switched rhs dataset. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 234–249.

Naome A Etori, Kevin Lu, Randu Karisa, and Arturs Kanepajs. 2025. Lag-mmlu: Benchmarking frontier llm understanding in latvian and giriama. *arXiv preprint arXiv:2503.11911*.

Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021. Many-to-english machine translation tools, data, and pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 306–316.

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, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jiaxin Guo, Hao Yang, Zongyao Li, Daimeng Wei, Hengchao Shang, and Xiaoyu Chen. 2024. A novel paradigm boosting translation capabilities of large language models. *arXiv preprint arXiv:2403.11430*.

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, et al. 2018. Effective parallel corpus mining using bilingual sentence embeddings. *arXiv preprint arXiv:1807.11906*.

Daniel Han, Michael Han, and Unsloth team. 2023. Unsloth.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International conference on machine learning*, pages 4411–4421. PMLR.

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

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. *arXiv preprint arXiv:1805.12282*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagnè, Alexandra Sasha Luccioni, François Yvon, and Matthias Gallé. 2023. Bloom: A 176b-parameter open-access multilingual language model. Preprint. Available at `https://inria.hal.science/hal-03850124/`.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2024. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. *Transactions of the Association for Computational Linguistics*, 12:576–592.

Edwin A. Locke and Gary P. Latham. 2002. Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, 57(9):705–717.

Laura Martinus and Jade Z. Abbott. 2019. A Focus on Neural Machine Translation for African Languages. ArXiv:1906.05685 [cs].

Audrey Mbogho, Quin Awuor, Andrew Kipkebut, Lilian Wanzare, and Vivian Oloo. 2025. Building low-resource african language corpora: A case study of kidawida, kalenjin and dholuo. arXiv preprint arXiv:2501.11003.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Peter Nabende, Naomi Muzaki, Claire Babirye, Jonathan Mukiibi, Jeremy Tusubira, Joyce Nakatumba-Nabende, and Andrew Katumba. 2023. Lumasaba Monolingual Corpus.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Hassan Muhammad, Salomon Kabongo, Salomey Osei, Sackey Freshia, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa, Mofe Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Jane Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkabir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Espoir Murhabazi, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Emezue, Bonaventure Dossou, Blessing Sibanda, Blessing Itoro Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. ArXiv:2010.02353 [cs].

Raymond Ng, Thanh Ngan Nguyen, Yuli Huang, Ngee Chia Tai, Wai Yi Leong, Wei Qi Leong, Xianbin Yong, Jian Gang Ngui, Yosephine Susanto, Nicholas Cheng, et al. 2025. Sea-lion: Southeast asian languages in one network. *arXiv preprint arXiv:2504.05747*.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, et al. 2023. Seallms–large language models for southeast asia. *arXiv preprint arXiv:2312.00738*.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2023. Afrobench: How good are large language models on african languages. In *18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1963–1978.

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.

Akintunde Oladipo, Mofetoluwa Adeyemi, Orevaoghene Ahia, Abraham Owodunni, Odunayo Ogundepo, David Adelani, and Jimmy Lin. 2023. Better quality pre-training data and t5 models for african languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 158–168.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan van Biljon, Arshath Ramkilowan, Adewale Akinfaderin, Alp Öktem, Wole Akin, Ghollah Kioko, Kevin Degila, Herman Kamper, Bonaventure Dossou, Chris Emezue, Kelechi Ogueji, and Abdallah Bashir. 2020a. Masakhane – Machine Translation For Africa. ArXiv:2003.11529 [cs].

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, et al. 2020b. Masakhane–machine translation for africa. *arXiv preprint arXiv:2003.11529*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, page 311, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all–adapting pre-training data processing to every language. *arXiv preprint arXiv:2506.20920*.

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

Matt Post, Martin Popel, and Ozan Caglayan. 2025. Sacrebleu. Version 2.5.1.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Machel Reid, Junjie Hu, Graham Neubig, and Yutaka Matsuo. 2021a. AfroMT: Pretraining strategies and reproducible benchmarks for translation of 8 african languages. arXiv preprint arXiv:2109.04715.

Machel Reid, Junjie Hu, Graham Neubig, and Yutaka Matsuo. 2021b. AfroMT: Pretraining Strategies and Reproducible Benchmarks for Translation of 8 African Languages. ArXiv:2109.04715 [cs].

Nathaniel R Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Onenamiyi Onesi, Claire Bizon Monroc, Loïc Grobol, Hasan Muhammad, Ashi Garg, Naome A Etori, et al. 2024. Krey\ol-mt: Building mt for latin american, caribbean and colonial african creole languages. *arXiv preprint arXiv:2405.05376*.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, et al. 2021. Xtreme-r: Towards more challenging and nuanced multilingual evaluation. *arXiv preprint arXiv:2104.07412*.

Felipe Sánchez-Martínez, Víctor M Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, Mikel L Forcada, Miquel Espla-Gomis, Andrew Secker, Susie Coleman, and Julie Wall. 2020. An english-swahili parallel corpus and its use for neural machine translation in the news domain. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 299–308.

Edoardo Signoroni and Pavel Rychlỳ. 2023. Evaluating sentence alignment methods in a low-resource setting: An english-yorùbá study case. In *Proceedings of the Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*, pages 123–129.

Steve Sloto, Brian Thompson, Huda Khayrallah, Tobias Domhan, Thamme Gowda, and Philipp Koehn. 2023. Findings of the wmt 2023 shared task on parallel data curation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 95–102.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team. 2025a. Gemma 3.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

NLLB Team, Marta R. Costa-jussà, James Cross andOnur Çelebiand Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang anf Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Qwen Team. 2025b. Qwen3 technical report.

Jörg Tiedemann. 2012. OPUS – an open source parallel corpus. In *Proceedings of LREC 2012*, pages 2214–2218.

Atnafu Lambebo Tonja, Bonaventure FP Dossou, Jessica Ojo, Jenalea Rajab, Fadel Thior, Eric Peter Wairagala, Anuoluwapo Aremu, Pelonomi Moiloa, Jade Abbott, Vukosi Marivate, et al. 2024a. Inkubalm: A small language model for low-resource african languages. *arXiv preprint arXiv:2408.17024*.

Atnafu Lambebo Tonja, Olga Kolesnikova, Alexander Gelbukh, and Jugal Kalita. 2024b. Ethiomt: Parallel corpus for low-resource ethiopian languages. *arXiv preprint arXiv:2403.19365*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 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.

Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, et al. 2023. Afrimte and africomet: Enhancing comet to embrace under-resourced african languages. *arXiv preprint arXiv:2311.09828*.

Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, et al. 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.

Barack Wanjawa, Lilian D.A. Wanzare, Florence Indede, Owen McOnyango, Edward Ombui, and Lawrence Muchemi. 2022a. Kencorpus: Kenyan languages corpus.

Barack Wanjawa, Lilian D.A. Wanzare, Florence Indede, Owen McOnyango, Edward Ombui, and Lawrence Muchemi. 2022b. Kencorpus: Kenyan Languages Corpus.

Lilian D.A Wanzare, Florence Indede, Owen McOnyango, Edward Ombui, Barack Wanjawa, and Lawrence Muchemi. 2022. KenTrans: A Parallel Corpora for Swahili and local Kenyan Languages.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Haoran Xu, Baolin Peng, Hany Awadalla, Dongdong Chen, Yen-Chun Chen, Mei Gao, Young Jin Kim, Yunsheng Li, Liliang Ren, Yelong Shen, et al. 2025. Phi-4-mini-reasoning: Exploring the limits of small reasoning language models in math. *arXiv preprint arXiv:2504.21233*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

A Literature Review

B Evaluation Benchmarks, Dataset Creation, and MT Resources

Multilingual evaluation benchmarks such as MEGA (Ahuja et al., 2023), XTREME (Hu et al., 2020), FLORES-101 (Goyal et al., 2022), FLORES-200 (Team et al., 2022), XTREME-R (Ruder et al., 2021), and COMET (Rei et al., 2020) cover only a small set of African languages, limiting their utility for assessing MT quality in low-resource settings. Yet an MT system is only as good as its training data (Sloto et al., 2023): high-quality parallel corpora remain the gold standard, providing the aligned bilingual signal required for robust translation learning (Signoroni and Rychlỳ, 2023; Sánchez-Martínez et al., 2020; Duh et al., 2020; Locke and Latham, 2002; Brown et al., 1990). Acquiring clean bitexts, however, is one of MT's most persistent challenges, especially for LRLs (Guo et al., 2018), where noise from misalignment, untranslated segments, or incorrect translations can significantly degrade performance (Khayrallah and Koehn, 2018). Parallel data are typically created through document alignment, sentence alignment, or extraction from comparable corpora (Sloto et al., 2023), and we leverage all these approaches to produce clean bitexts (Signoroni and Rychlỳ, 2023). Several initiatives have advanced African MT—including Masakhane (Orife et al., 2020b), MAFAND-MT (Adelani et al., 2022b), AfroBench (Ojo et al., 2025), Cheetah (Adebara et al., 2024), MMTAfrica (Emezue and Dossou, 2022b), and EthioMT (Tonja et al., 2024b). AfriMMT-EA extends this line of work by introducing large-scale, multi-domain parallel corpora for 53 East African languages.

C Domain Composition

C.1 Uganda

Uganda. Ugandan languages form the second largest share of the corpus (35.5%; Table 2). Most data come from *UgandaLex2*[7], a Bible-based parallel dataset for 24 languages, complemented by smaller web and community contributions. We further expand domain coverage using Luganda resources, such as *Luganda010224*, *Luganda_Sci-Math-Bio_Translations*, *Crowd-Validated-Paths*, and *luganda_test_select*—covering educational, scientific, and community-curated (Appendix J

C.2 Ethiopia and Rwanda

Ethiopia and **Rwanda** regions contributes 0.5% and 0.9% of the corpus respectively, shown in Table 2, primarily sourced from publicly available published materials, composed of mixed [11]and religious domain . **Table 7** details the distribution of translation pairs, token counts, and language representation, while **Table 3** provides ISO codes, language families, regions spoken for each language.

D Language and Translation Directions

AfriMMT-EA covers **53** languages from five East African countries: **24** in Kenya, **15** in Uganda, **11** in Tanzania, **1** in Rwanda, and **2** in Ethiopia. These languages fall within the Niger–Congo, Nilo–Saharan, and Afro–Asiatic families. The benchmark includes **11** Tanzanian and **12** Kenyan languages with no prior MT coverage. Swahili and English serve as pivot languages (Table 2). We provide **66** translation directions: **19 English→local** and **47 Swahili→local**. Of these, **13** languages include both English→X and Swahili→X (26 directions). The remaining **6 English only** and **34 Swahili only** directions show available data (Table 11).

E Data Format

Table 6 presents the CSV schema used for the aligned dataset. The collection contains a total of 714,490 translation pairs, each stored as a structured row specifying the source and target languages, the original text, its tokenized representation, the translated segment, and the corresponding translation token. The schema also includes country information to support geographic attribution and dialect sensitive analysis. Metadata fields such as the processing timestamp and the pipeline version provide full traceability of data creation and enable reproducibility across preprocessing and alignment stages.

Quality metrics computed over the full dataset show that the average source text length is 86.1 characters, and the average translation length is 90.5 characters, reflecting the overall balance and consistency of the aligned segments. This standardized format facilitates reliable parsing, multilingual analyses, and seamless integration with downstream training and evaluation pipelines.

[11] https://github.com/AAUThematic4LT

F Argilla Web-based Translation Interface

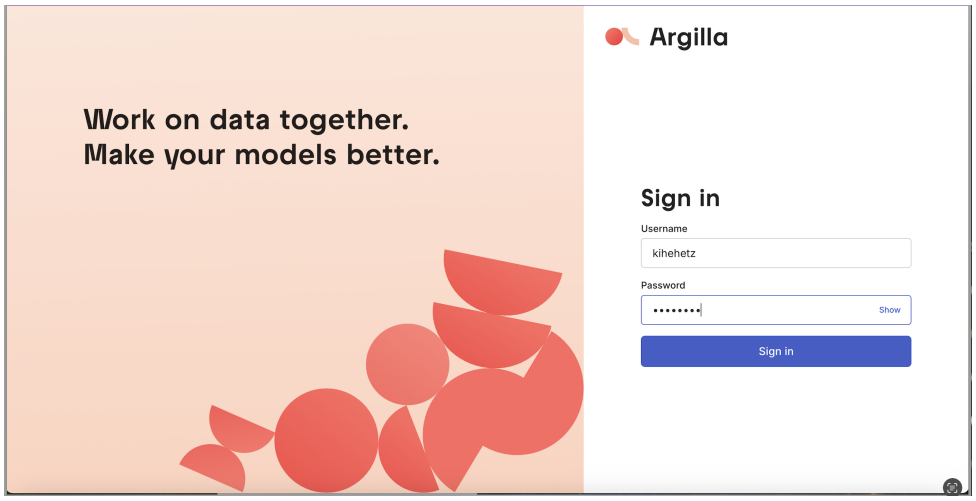

Figure 6: Argilla web-based translation interface. Each contributor is provided with a unique username and password for the specific language they wish to contribute to, allowing controlled access to submit translation data.

G Instruction to Volunteers

H Gemini Data Extraction Interface

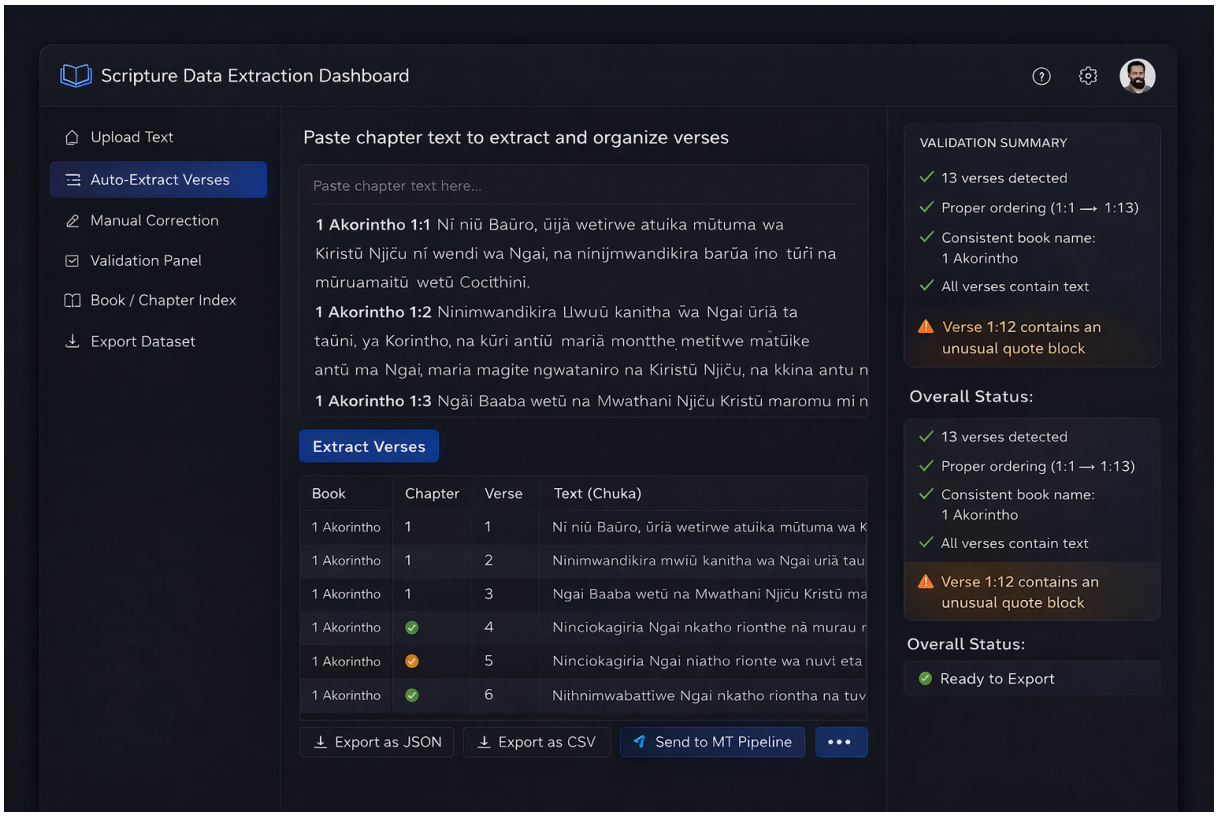

Figure 9: **Gemini Data Extraction Interface.** Internal interface used to automatically segment text data or bible chapter text into verse-aligned or parallel units, validate ordering and consistency, and export structured data for downstream annotation and MT pipelines.

I Prompt for Each Model

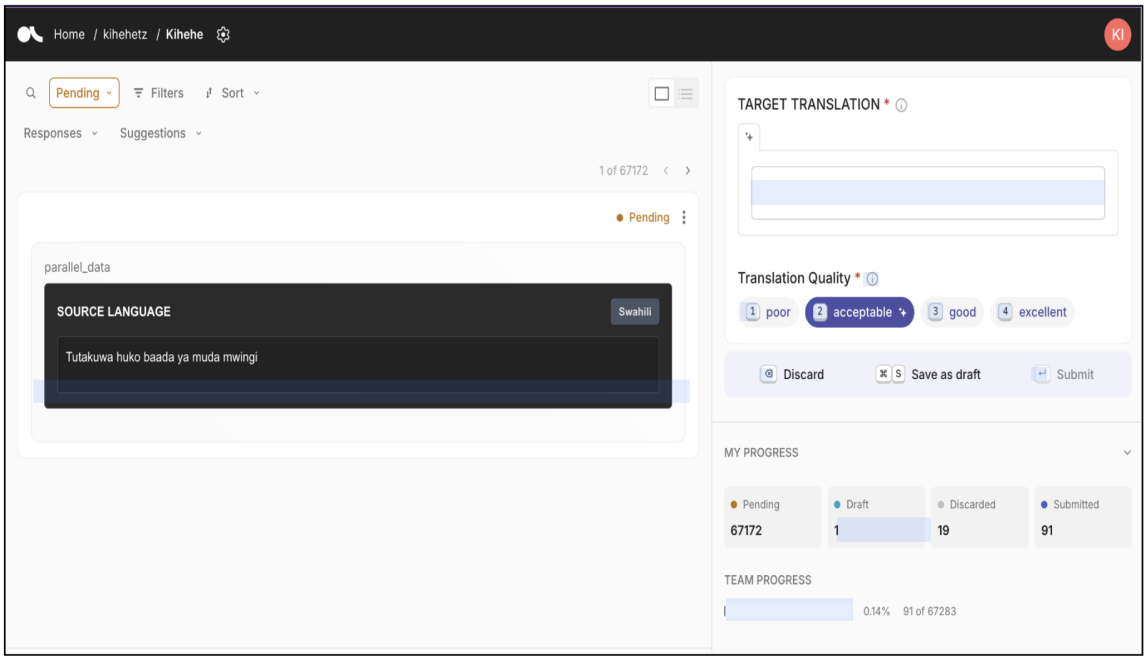

Figure 7: Annotation interface for Kihehe parallel data showing the Swahili source sentence and target translation input workflow. This is what a translator sees after logging in the Argilla web-based interface shown in Figure 6 .

Listing 1: ChatML prompt template used for training and evaluation.

```
def convert_to_chatml(example):
    input_text = (
        f"translate from {example['source_language']} to "
        f"{example['translated_language']}\n"
        f"{example['source']}"
    )
    return {
        "conversations": [
            {
                "role": "system",
                "content": (
                    "You are a translator assistant that "
                    "translates text from one language to another."
                )
            },
            {
                "role": "user",
                "content": input_text
            },
            {
                "role": "assistant",
                "content": example['translation']
            }
        ]
    }
```

J External Resources and Dataset URLs

K Data Format

Data Format: Our aligned datasets were stored in CSV format using a schema shown in Table 6 below. Each record represents a single translation pair with metadata describing its source, processing stage, and origin.

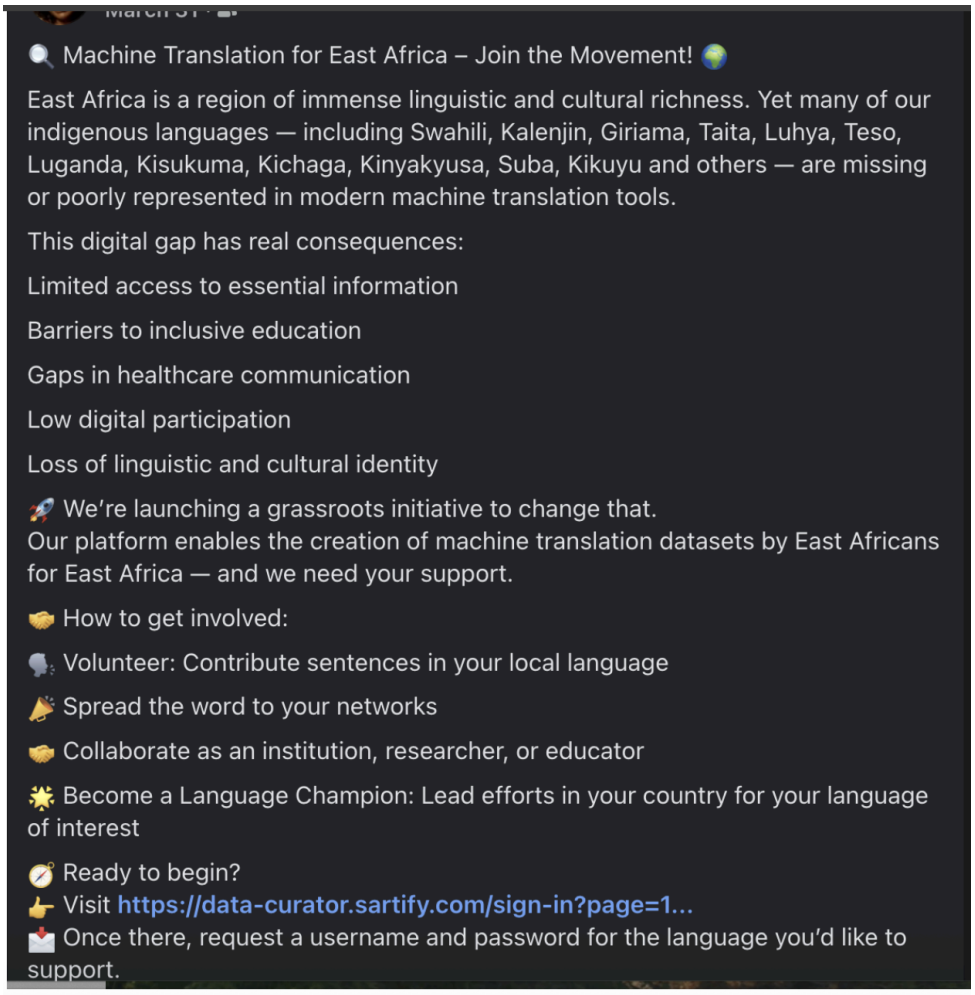

Figure 8: Community engagement post announcing the AfriMMT-EA initiative. The campaign invited volunteers, researchers, and institutions across East Africa to contribute translation data for low-resource languages, promoting inclusive participation and grassroots dataset creation.

L Language Distribution

As shown in Table 7 Our datasets comprises translation pairs across 53 East African languages from five countries: Uganda, Kenya, Tanzania, Rwanda, and Ethiopia. Table 7 presents the distribution of translation pairs, token counts, and language representation across the dataset. The distribution of languages in the dataset is uneven. Luganda (Uganda) dominates, accounting for 22.7% of the corpus with 162,446 translation pairs and over 8 million tokens. Kikuyu (Kenya) is the second most represented language at 14.0%, followed by Taita (Kenya) at 8.2% and English (Kenya) at 7.6%. Collectively, these four languages constitute more than half of the dataset.

The majority of languages (37 out of 52) represent less than 1% of the dataset each, highlighting the long-tail distribution typical of low-resource language collections. Kenya contributes the most languages to the dataset (21 languages), followed by Uganda (17 languages) and Tanzania (11 languages). Rwanda and Ethiopia contribute one and two languages respectively. Token counts vary significantly across languages, with source tokens ranging from 408 (Tigrinya) to 3,079,301 (Luganda), and translation tokens ranging from 90 (Tigrinya) to 5,061,186 (Luganda). This variation reflects both the different dataset sizes and the morphological characteristics of each language.

Name	URL
Bible Society Kenya	btlkenya.org
Bible.com	bible.com
Google Gemini	gemini.google.com
UgandaLex2	UgandaLex2
Luganda010224	Luganda010224
Luganda_Sci-Math-Bio_Translations	Luganda Sci–Math–Bio
Crowd-Validated-Paths	Luganda Crowd Paths
Data Curator Portal	data-curator.sartify.com
Ethiopian Parallel Corpora	Ethiopian Corpora
ACL Anthology	aclanthology.org
Scholar Search	scholar.google.com
Lanfrica	lanfrica.com
Google Search	google.com
Gemini	gemini.google.com
Archive.org	archive.org
Facebook	facebook.com
LinkedIn	linkedin.com

Table 5: External resources and dataset URLs referenced in this work.

Column	Type	Description
`source_language`	string	The source language (typically Swahili)
`translated_language`	string	The target language
`source`	string	The source text
`source_token`	string	The source token of the original text
`translation_token`	string	The translation token of the original text
`translation`	string	The translated text
`country`	string	The country of origin
`created_date`	datetime	Processing timestamp
`processor_version`	string	Version of the processing pipeline

Table 6: Schema of the aligned dataset stored in CSV format.

M Experiments Results

N Fertility Analysis and Sequence Length Distribution

To characterize structural variation across the languages in our corpus, we compute two standard corpus-level statistics used in machine translation: *average sequence length* and *fertility*. For each language ℓ, given N_ℓ translation pairs with S_ℓ total source tokens and T_ℓ total target tokens, we define:

$$\text{AvgSeqLen}_{\text{src}}(\ell) = \frac{S_\ell}{N_\ell}, \qquad \text{AvgSeqLen}_{\text{tgt}}(\ell) = \frac{T_\ell}{N_\ell},$$

$$\text{Fertility}(\ell) = \frac{T_\ell}{S_\ell}.$$

Average sequence lengths quantify the typical sentence or segment length associated with each language, while fertility captures the degree of translation length expansion. High fertility values (> 1) indicate that target translations tend to be longer than their corresponding sources, often reflecting richer inflectional or agglutinative morphology. Lower fertility values (< 1) suggest more analytic structures or more compact surface forms.

Across the corpus, we observe substantial variation in both metrics. Many Bantu languages (e.g., Luganda, Kikuyu, Suba, Zanaki) exhibit fertility values in the 1.2–1.6 range and longer average target sequences, consistent with their complex agreement morphology and noun-class systems. Nilotic languages such as Luo, Pokot and English pairs show fertility closer to 1.0 or below, reflecting more analytic morphology. Tanzanian Bantu languages such as Gogo, Makonde, and Zanaki show some of the longest sequences in the dataset (40–60 tokens on average), whereas languages like Taita, Samburu, and Chonyi exhibit much shorter average lengths (< 15 tokens). The results highlight the significant typological diversity of East African languages and underscore the importance of adapting MT models to handle

Table 7: Language Statistics by Country

Language	Country	Translation Pairs	Source Tokens	Translation Tokens	Total Tokens	Percentage of language
Luganda	Uganda	162,446	3,079,301	5,061,186	8,140,487	22.7%
Kikuyu	Kenya	100,195	2,141,717	3,510,370	5,652,087	14.0%
Taita	Kenya	58,537	809,231	888,257	1,697,488	8.2%
English	Kenya	54,221	1,020,199	739,610	1,759,809	7.6%
Gogo	Tanzania	32,371	1,456,656	1,865,915	3,322,571	4.5%
Luhya	Kenya	21,560	317,414	339,491	656,905	3.0%
Chaga	Tanzania	10,106	351,653	393,654	745,307	1.4%
Luo	Kenya	10,038	362,766	356,761	719,527	1.4%
Borana	Kenya	9,697	247,267	264,151	511,418	1.4%
Giriama	Kenya	9,629	327,412	376,830	704,242	1.3%
Makonde	Tanzania	9,111	395,038	516,338	911,376	1.3%
Jita	Tanzania	8,820	333,545	361,195	694,740	1.2%
Bena	Tanzania	8,674	334,552	480,999	815,551	1.2%
Kamba	Kenya	8,210	212,480	220,388	432,868	1.1%
Kalenjin	Kenya	7,969	84,952	135,554	220,506	1.1%
Suba	Kenya	7,965	361,159	437,423	798,582	1.1%
Hehe	Tanzania	7,849	325,353	365,460	690,813	1.1%
Langi	Tanzania	7,849	322,965	409,292	732,257	1.1%
Zanaki	Tanzania	7,813	321,367	485,813	807,180	1.1%
Pokot	Kenya	6,533	234,343	217,501	451,844	0.9%
Gusii	Kenya	6,494	269,601	261,580	531,181	0.9%
Sukuma	Tanzania	6,394	192,310	227,498	419,808	0.9%
Masaba	Uganda	6,246	231,741	328,161	559,902	0.9%
Acholi	Uganda	6,246	231,740	242,244	473,984	0.9%
Alur	Uganda	6,246	231,740	231,136	462,876	0.9%
Ateso	Uganda	6,246	231,740	280,299	512,039	0.9%
Kakwa	Uganda	6,246	231,740	289,423	521,163	0.9%
Gwere	Uganda	6,246	231,740	340,491	572,231	0.9%
Jopadhola	Uganda	6,246	231,740	247,727	479,467	0.9%
Lango	Uganda	6,246	231,741	237,077	468,818	0.9%
Kebu	Uganda	6,246	231,740	265,804	497,544	0.9%
Kinyarwanda	Rwanda	6,246	231,740	235,655	467,395	0.9%
Nyole	Uganda	6,246	231,740	294,609	526,349	0.9%
Nyoro	Uganda	6,246	231,740	246,490	478,230	0.9%
Soga	Uganda	6,246	231,740	267,777	499,517	0.9%
Kumam	Uganda	6,246	231,740	265,015	496,755	0.9%
Aringa	Uganda	6,243	231,605	229,863	461,468	0.9%
Meru	Kenya	5,420	207,053	206,282	413,335	0.8%
Samburu	Kenya	5,411	47,140	45,595	92,735	0.8%
Embu	Kenya	5,349	228,901	226,948	455,849	0.7%
Maasai	Tanzania	5,327	154,089	190,289	344,378	0.7%
Taveta	Kenya	5,249	228,860	186,804	415,664	0.7%
Marakwet	Kenya	5,249	228,624	266,270	494,894	0.7%
Chuka	Kenya	5,030	205,331	195,227	400,558	0.7%
Duruma	Kenya	5,011	197,413	248,717	446,130	0.7%
Somali	Kenya	5,011	197,413	235,446	432,859	0.7%
Pokomo	Kenya	5,010	197,338	246,456	443,794	0.7%
Oromo	Ethiopia	3,423	77,261	160,014	237,275	0.5%
Haya	Tanzania	1,306	19,864	21,639	41,503	0.2%
Pare	Tanzania	1,000	15,001	18,819	33,820	0.1%
Chonyi	Kenya	370	1,359	1,291	2,650	0.1%
Swahili	Kenya	138	2,668	4,200	6,868	0.0%
Tigrinya	Ethiopia	18	408	90	498	0.0%

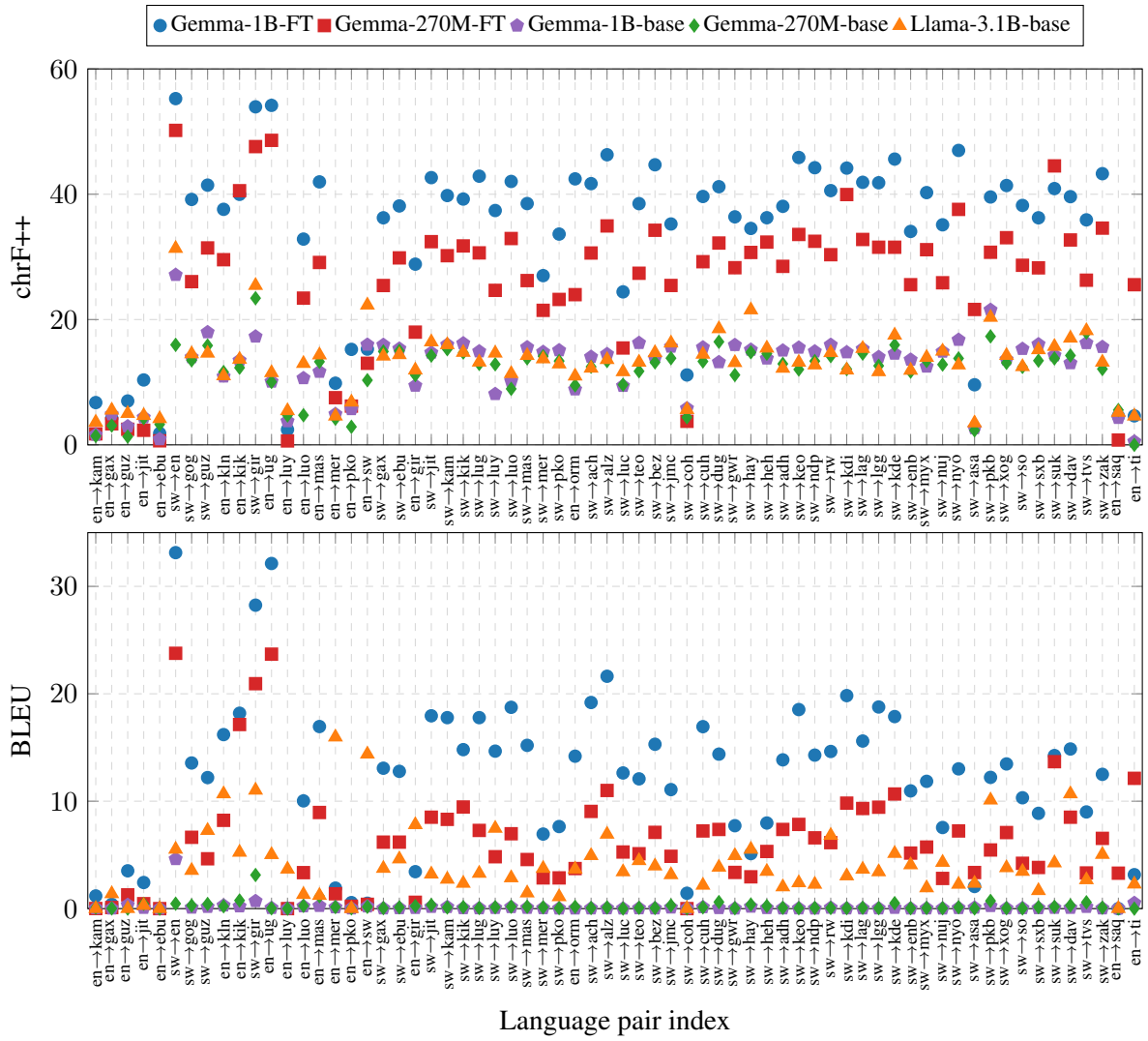

Figure 10: chrF++and BLEU scores for English→local and Swahili→local across 66 language pairs for models

wide variation in morphological richness, sequence length, and translation length divergence As shown in Table 12.

O Open Models Analysis

P Instruction to Translators using Argilla Platform

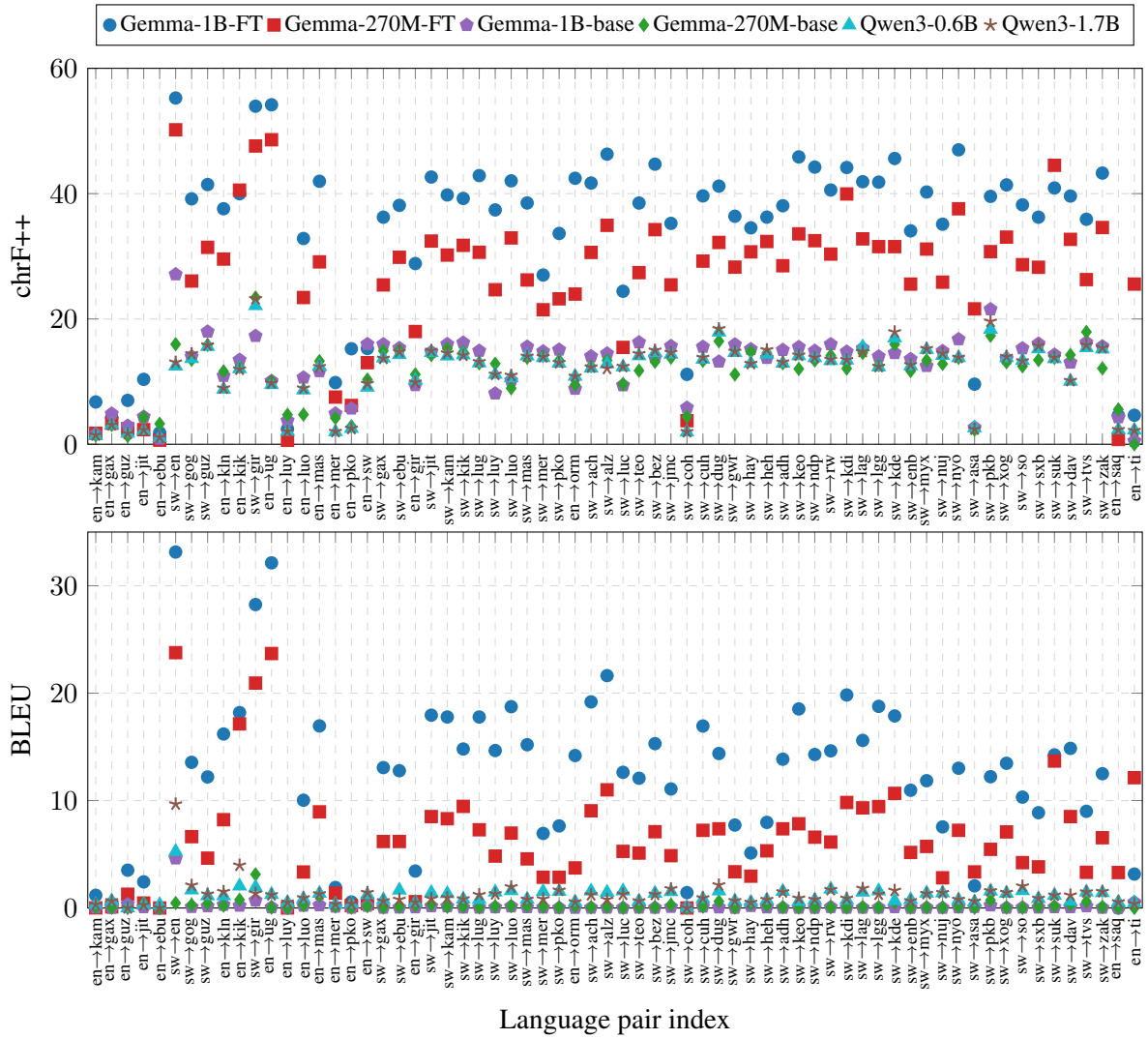

Figure 11: chrF++and BLEU scores for English→local and Swahili→local across 66 language pairs for models.

Name	Country	Translation pairs	Old Tokens	New Tokens	Domain	source
Swahili	Kenya	138	2,668	-	Short stories, Politics, Healthcare	Publications
Duruma	Kenya	5,011	-	248,717	Religion	Bible
Suba	Kenya	7,965	-	437,423	Religion	Bible
Taita	Kenya	58,537 (new: 5,219)	802,080	86,177	Religion, Mixed domain (news, education, government, politics, healthcare)	Bible, Publications (Mbogho et al., 2025)
Meru	Kenya	5,420	-	206,282	Religion	Bible
Gusii	Kenya	6,494	-	261,580	Religion, Lexical, Mixed domain (news, education, government, politics, healthcare)	Bible, Publications (Mbogho et al., 2025)
Saamia	Kenya	107	-	289	Lexical	Web Articles
Kamba	Kenya	8,210	-	220,388	Lexical, Religion, Mixed domain	Bible, Web Articles
Teso	Kenya	899	-	1,467	Lexical (Wordlist)	Web Articles (Internet Archive)
Kalenjin	Kenya	7,969	-	7,969	Lexical, Healthcare, Survey, Mixed domain (news, education, government, politics, healthcare)	Web Articles, Private Data
Embu	Kenya	5,349	-	226,948	Religion	Bible
Chonyi	Kenya	370	-	1,291	Lexical	Web Articles
Luwanga	Kenya	7,337	-	7,736	Dictionary	Web Articles
Kikuyu	Kenya	100,195 (new pairs: 5,011)	3,419,801	90,569	Agriculture, Religion, Lexical	NLLBv1 dataset, OPUS, MT560, CLICS, Bible
Luo	Kenya	10,038	349,277	7,484	Children's Literature / African Literature, Mixed domain (news, education, government, politics, healthcare)	Children's Literature / African Literature
Maasai	Kenya	4,375	-	61,863	Religion, African literature, Lexical, Mixed domain (news, education, government, politics, healthcare)	Bible, Translation of Existing Publications (Mbogho et al., 2025), Web Articles
Borana	Kenya	9,697	-	264,151	Religion	Bible, Dictionary
Tharaka	Kenya	5,249	-	92,865	Religion	Bible
Bukusu	Kenya	4,240	-	22,311	Dictionary, Religious	Dictionary, Religion
Maragoli	Kenya	9,146 (new: 5,678)	23,445	35,632	Cultural Narrative (Folkloric and Oral Tradition), Conversational Folklore	Publications - KenCorpus, Web Articles
Chuka	Kenya	5,030	-	195,227	Religion	Bible
Samburu	Kenya	5,411	-	45,595	Dictionary	Dictionary
Marakwet	Kenya	5,249	-	266,270	Religion	Bible
Taveta	Kenya	5,249	-	186,804	Religion	Bible
Giriama	Kenya	9,629	-	376,830	Religion, Cultural Narrative, Oral Literature, Cultural Heritage, Mixed domain (news, education, government, law, politics, healthcare, political, public health)	Bible, Translation of Existing Publications (KenCorpus), Web Articles, Book Translation
Pokomo	Kenya	5,010	-	246,456	Religious	Bible
Pokot	Kenya	6,533	-	217,501	Religious	Bible
Somali	Kenya	5,011	-	235,446	Religious	Bible
Oromo	Ethiopia	3,423	160,014	-	Mixed	Published resources
Tigrinya	Ethiopia	18	90	-	Mixed	Published resources
Kinyarwanda	Rwanda	6,246	235,655	-	Religion	Bible
Bena	Tanzania	8,674	-	480,999	Religion, Mixed Domains	Bible and Published resources
Gogo	Tanzania	32,371	-	1,865,915	Religion	Bible
Makonde	Tanzania	9,111	-	516,338	Religion	Bible
Langi	Tanzania	7,849	-	409,292	Religion	Bible
Maasai	Tanzania	5,327	-	190,289	Religion	Bible
Jita	Tanzania	8,820	-	361,195	Religion, Mixed Domains	Bible and Published resources
Kihehe	Tanzania	7,849	-	365,460	Religion, Mixed Domains	Bible and Published resources
Kichaga	Tanzania	10,106	-	393,654	Religion, Mixed Domains	Bible and Published resources
Kisukuma	Tanzania	6,394	-	227,498	Literature, Mixed Domains	Published resources
Kihaya	Tanzania	1,306	-	21,639	Religion, Mixed Domains	Bible and Published resources
Pare	Tanzania	1,000	-	18,819	Mixed Domains	Bible
Acholi	Uganda	6,246	231,740	-	Religion, Mixed Domains	Bible and Published resources
Alur	Uganda	6,246	231,740	-	Religion	Bible
Aringa	Uganda	6,243	231,605	-	Religion	Bible
Ateso	Uganda	6,246	231,740	-	Religion	Bible
Gwere	Uganda	6,246	231,740	-	Religion	Bible
Jopadhola	Uganda	6,246	231,740	-	Religion	Bible
Kakwa	Uganda	6,246	231,740	-	Religion	Bible
Kebu	Uganda	6,246	231,740	-	Religion	Bible
Kumam	Uganda	6,246	231,740	-	Religion	Bible
Lango	Uganda	6,246	231,741	-	Religion	Bible
Lugbara	Uganda	162,446	3,079,301	-	Religion, Mixed Domains	Bible and Published resources
Masaaba	Uganda	6,246	231,741	-	Religion, Mixed Domains	Bible and Published resources
Nyole	Uganda	6,246	231,740	-	Religion	Bible
Nyoro	Uganda	6,246	231,740	-	Religion	Bible
Soga	Uganda	6,246	231,740	-	Religion	Bible

Table 8: Specific token and translation pair contributions for the 53 East African Languages considered

Name	Country	Translation pairs	Old Tokens	New Tokens	Domain	source
Swahili	Kenya	138	2,668	-	Short stories, Politics, Healthcare	Publications
Duruma	Kenya	5,011	-	**248,717**	Religion	Bible
Suba	Kenya	7,965	-	**437,423**	Religion	Bible
Taita	Kenya	58,537 (new: 5,219)	802,080	**86,177**	Religion, Mixed domain (news, education, government, politics, healthcare)	Bible, Publications (Mbogho et al., 2025)
Meru	Kenya	5,420	-	**206,282**	Religion	Bible
Gusii	Kenya	6,494	-	**261,580**	Religion, Lexical, Mixed domain	Bible, Publications (Mbogho et al., 2025)
Saamia	Kenya	107	-	**289**	Lexical	Web Articles
Kamba	Kenya	8,210	-	**220,388**	Lexical, Religion, Mixed domain	Bible, Web Articles
Teso	Kenya	899	-	**1,467**	Lexical (Wordlist)	Web Articles (Internet Archive)
Kalenjin	Kenya	7,969	-	**7,969**	Lexical, Healthcare, Survey, Mixed domain	Web Articles, Private Data
Embu	Kenya	5,349	-	**226,948**	Religion	Bible
Chonyi	Kenya	370	-	**1,291**	Lexical	Web Articles
Luwanga	Kenya	7,337	-	**7,736**	Dictionary	Web Articles
Kikuyu	Kenya	100,195 (new: 5,011)	3,419,801	**90,569**	Agriculture, Religion, Lexical	NLLBv1 dataset, OPUS, MT560, CLICS, Bible
Luo	Kenya	10,038	349,277	**7,484**	Literature, Mixed domain	Children's Literature
Maasai	Kenya	4,375	-	**61,863**	Religion, African literature, Lexical	Bible, Publications
Borana	Kenya	9,697	-	**264,151**	Religion	Bible, Dictionary
Tharaka	Kenya	5,249	-	**92,865**	Religion	Bible
Bukusu	Kenya	4,240	-	**22,311**	Dictionary, Religious	Dictionary, Religion
Maragoli	Kenya	9,146 (new: 5,678)	23,445	**35,632**	Cultural Narrative, Folklore	KenCorpus, Web Articles
Chuka	Kenya	5,030	-	**195,227**	Religion	Bible
Samburu	Kenya	5,411	-	**45,595**	Dictionary	Dictionary
Marakwet	Kenya	5,249	-	**266,270**	Religion	Bible
Taveta	Kenya	5,249	-	**186,804**	Religion	Bible
Giriama	Kenya	9,629	-	**376,830**	Religion, Cultural Narrative, Oral Lit	Bible, KenCorpus, Web Articles
Pokomo	Kenya	5,010	-	**246,456**	Religious	Bible
Pokot	Kenya	6,533	-	**217,501**	Religious	Bible
Somali	Kenya	5,011	-	**235,446**	Religious	Bible
Oromo	Ethiopia	3,423	160,014	-	Mixed	Published resources
Tigrinya	Ethiopia	18	90	-	Mixed	Published resources
Kinyarwanda	Rwanda	6,246	235,655	-	Religion	Bible
Bena	Tanzania	8,674	-	**480,999**	Religion, Mixed Domains	Bible and Publications
Gogo	Tanzania	32,371	-	**1,865,915**	Religion	Bible
Makonde	Tanzania	9,111	-	**516,338**	Religion	Bible
Langi	Tanzania	7,849	-	**409,292**	Religion	Bible
Maasai	Tanzania	5,327	-	**190,289**	Religion	Bible
Jita	Tanzania	8,820	-	**361,195**	Religion, Mixed Domains	Bible and Publications
Kihehe	Tanzania	7,849	-	**365,460**	Religion, Mixed Domains	Bible and Publications
Kichaga	Tanzania	10,106	-	**393,654**	Religion, Mixed Domains	Bible and Publications
Kisukuma	Tanzania	6,394	-	**227,498**	Literature, Mixed Domains	Published resources
Kihaya	Tanzania	1,306	-	**21,639**	Religion, Mixed Domains	Bible and Publications
Pare	Tanzania	1,000	-	**18,819**	Mixed Domains	Bible
Acholi	Uganda	6,246	231,740	-	Religion, Mixed Domains	Bible and Publications
Alur	Uganda	6,246	231,740	-	Religion	Bible
Aringa	Uganda	6,243	231,605	-	Religion	Bible
Ateso	Uganda	6,246	231,740	-	Religion	Bible
Gwere	Uganda	6,246	231,740	-	Religion	Bible
Jopadhola	Uganda	6,246	231,740	-	Religion	Bible
Kakwa	Uganda	6,246	231,740	-	Religion	Bible
Kebu	Uganda	6,246	231,740	-	Religion	Bible
Kumam	Uganda	6,246	231,740	-	Religion	Bible
Lango	Uganda	6,246	231,741	-	Religion	Bible
Lugbara	Uganda	162,446	3,079,301	-	Religion, Mixed Domains	Bible and Publications
Masaaba	Uganda	6,246	231,741	-	Religion, Mixed Domains	Bible and Publications
Nyole	Uganda	6,246	231,740	-	Religion	Bible
Nyoro	Uganda	6,246	231,740	-	Religion	Bible
Soga	Uganda	6,246	231,740	-	Religion	Bible

Table 9: Specific token and translation pair contributions for the 53 East African Languages considered. The **New Tokens** column highlights, in **bold**, **our newly collected and curated datasets contributed by this paper**, while the ***Old Tokens*** column reflects previously available resources.

Pair Type	Lang Pair	n_samples	BLEU_270M	chrF++_270M	ter_270M	BLEU_1B	chrF++_1B	ter_1B
	english→borana	889	0.10	3.36	**103.98**	**0.40**	**4.25**	118.48
	english→embu	20	0	0.65	**100**	0	**1.83**	105
	english→giriama	26	0.59	17.99	164.76	**3.44**	**25.84**	**111.11**
	english→gusii	20	1.28	2.49	**104**	**3.52**	**7**	124
	english→jita	21	0.46	2.33	**102.56**	**2.43**	**10.36**	117.95
	english→kamba	180	0.01	1.76	**101.22**	**1.18**	**6.75**	105.69
	english→kikuyu	18989	17.14	**40.56**	81.97	**18.19**	39.98	**79.45**
	english→luganda	23875	23.69	48.60	77.47	**32.13**	**54.18**	**69.68**
	english→luhya	1518	0	0.63	**100.12**	**0.03**	**2.43**	103.38
	english→luo	154	3.35	23.42	129.76	**10.03**	**32.84**	**93.06**
	english→maasai	416	8.95	29.11	106.15	**16.95**	**41.97**	**82.29**
	english→meru	39	1.38	7.53	**107.46**	**1.91**	**9.84**	152.46
	english→pokot	260	0.22	6.19	**123.4**	**0.56**	**15.24**	106.60
	english→swahili	26	0.40	13.02	203.32	**0.50**	**15.88**	**143.21**
English and Swahili Paired	swahili→borana	1044	6.19	25.43	**114.37**	**13.07**	**36.24**	**92.67**
	swahili→embu	1044	6.19	29.84	114.47	**12.78**	**38.13**	**87.15**
	swahili→giriama	1885	20.94	47.60	77.75	**28.25**	**53.95**	**67.04**
	swahili→gusii	1269	4.64	31.43	125.30	**12.20**	**41.45**	**89.87**
	swahili→jita	1742	8.51	32.43	106.28	**17.95**	**42.65**	**85.85**
	swahili→kamba	1033	8.31	30.19	108.60	**17.78**	**39.79**	**85.26**
	swahili→kikuyu	1002	9.46	31.75	99.03	**14.80**	**39.23**	**84.89**
	swahili→luganda	1248	7.28	30.64	107.87	**17.78**	**42.88**	**75.86**
	swahili→luhya	2164	4.83	24.66	123.65	**14.66**	**37.40**	**88.59**
	swahili→luo	1577	6.97	32.93	99.26	**18.94**	**42.04**	**68.01**
	swahili→maasai	643	4.57	26.62	126.50	**15.20**	**38.51**	**87.93**
	swahili→meru	1044	2.86	21.47	150.08	**6.94**	**27.01**	**100.21**
	swahili→pokot	1044	2.86	23.21	129.80	**7.64**	**33.04**	**95.76**
	swahili→english	10844	23.77	50.19	67.53	**33.14**	**55.26**	**60.28**
	english→kalenjin	1431	8.22	29.55	127.90	**16.20**	**37.59**	**91.61**
	english→oromo	528	3.73	23.97	127.24	**14.19**	**42.45**	**88.03**
English Paired Only	english→samburu	1073	3.30	4.59	**100.10**	0	**0.77**	102.11
	english→tigrinya	4	**12.13**	**25.56**	**150**	3.16	4.63	300
	swahili→acholi	1248	9.05	30.62	102.02	**19.19**	**41.69**	**78.84**
	swahili→alur	1249	11.00	34.94	94.73	**21.06**	**46.30**	**74.69**
	swahili→aringa	1247	5.27	15.46	103	**12.63**	**24.42**	**84.14**
	swahili→ateso	1249	5.11	27.39	116.06	**12.08**	**38.50**	**88.29**
	swahili→bena	1731	7.10	34.25	108.13	**15.30**	**44.70**	**85.76**
	swahili→chaga	1949	4.87	25.44	134.38	**11.08**	**35.25**	**97.58**
	swahili→chonyi	73	0	3.76	**102.70**	**1.43**	**11.16**	117.57
	swahili→chuka	951	7.23	29.23	112.43	**16.94**	**39.64**	**87.47**
	swahili→duruma	1001	7.37	32.21	108.46	**14.38**	**41.20**	**88.12**
	swahili→ganda	1247	11.53	34.51	93.53	**16.44**	**40.11**	**88.07**
	swahili→gogo	6451	6.64	26.05	114.30	**13.56**	**39.17**	**87.61**
	swahili→gwere	1248	3.37	28.27	122.51	**7.72**	**36.39**	**97.01**
	swahili→haya	247	2.96	30.71	94.88	**5.12**	34.54	**87.90**
	swahili→hehe	1570	5.32	32.36	115.94	**7.97**	**36.25**	**94.60**
	swahili→jopadhola	1249	7.37	28.49	99.41	**13.85**	**38.07**	**84.64**
	swahili→kakwa	1249	7.84	33.58	103.70	**18.53**	**45.86**	**76.66**
	swahili→kebu	1248	6.59	32.49	106.31	**14.29**	**44.23**	**80.85**
	swahili→kinyarwanda	1247	6.14	30.35	113.14	**14.63**	**40.58**	**88.86**
Swahili Paired Only	swahili→kumam	1248	9.92	33.95	99.44	**19.83**	**44.17**	**76.12**
	swahili→langi	1568	9.32	32.78	99.60	**15.60**	**41.90**	**85.53**
	swahili→lango	1248	9.44	31.54	94.23	**18.77**	**41.83**	**78.09**
	swahili→makonde	1820	10.67	36.77	104.98	**17.88**	**45.61**	**81.72**
	swahili→marakwet	1044	5.17	25.56	115.81	**10.96**	**34.06**	**97.07**
	swahili→masaba	1249	5.73	31.15	105.25	**11.85**	**40.26**	**88.79**
	swahili→nyole	1249	2.81	25.87	143.33	**7.55**	**35.12**	**99.61**
	swahili→nyoro	1249	7.23	37.58	99.99	**13.01**	46.99	**84.22**
	swahili→pare	198	3.36	21.61	135.18	**2.06**	9.59	**90.43**
	swahili→pokomo	1001	5.46	30.74	122	**12.22**	**39.56**	**90.27**
	swahili→soga	1248	7.08	33.06	99.87	**13.47**	**41.39**	**88.52**
	swahili→somali	1000	4.22	28.66	103.67	**10.32**	**38.21**	**87.97**
	swahili→suba	1587	3.83	28.25	120.40	**8.87**	**36.24**	**95.41**
	swahili→sukuma	1277	13.68	**44.52**	93.45	**14.25**	40.90	**86.43**
	swahili→taita	5438	8.51	32.70	105.07	**14.86**	**39.61**	**86.49**
	swahili→taveta	1044	3.32	26.38	137.35	**9**	**35.91**	**96.04**
	swahili→zanaki	1561	6.54	34.59	101.81	**12.50**	**43.28**	**91.02**
Averaged Summary		1918.43	6.47	26.35	111.78	**12.02**	**33.64**	**94.86**

Table 10: chrF++, BLEU, and TER scores for fine-tuned models in 66 East African language pairs across different model sizes. Higher is better for BLEU/chrF++; lower is better for TER.

Pair Type	Lang Pair	n_samples	BLEU_270M	chrF++_270M	TER_270M	BLEU_1B	chrF++_1B	TER_1B
	english→borana	889	**0.02**	3.08	**297.23**	0.01	**4.88**	1546.49
	english→embu	20	**0.00**	**3.27**	795.00	0.00	0.90	**210.00**
	english→giriama	26	**0.24**	**11.15**	**132.38**	0.06	9.44	366.98
	english→gusii	20	0.00	1.32	**152.00**	**0.34**	**2.95**	344.00
	english→jita	21	**0.44**	4.28	**238.46**	0.07	**4.38**	876.92
	english→kamba	180	**0.09**	1.42	**224.80**	0.04	**1.60**	334.15
	english→kikuyu	18989	**0.76**	12.31	**138.59**	0.21	**13.47**	259.40
	english→luganda	23875	0.03	**10.50**	**181.96**	**0.06**	10.09	303.27
	english→luhya	1518	**0.00**	**4.70**	**861.41**	0.00	3.77	2001.21
	english→luo	154	**0.26**	4.73	**122.53**	0.21	**10.64**	352.93
	english→maasai	416	**1.16**	**13.27**	**144.71**	0.25	11.66	306.70
	english→meru	39	**0.15**	4.18	**381.97**	0.12	**4.90**	549.18
	english→pokot	260	**0.03**	2.88	**290.94**	0.02	**5.70**	815.09
	english→swahili	26	0.17	10.32	**147.65**	**0.20**	**15.98**	303.88
English and Swahili	swahili→borana	1044	**0.05**	14.90	**145.22**	0.02	**15.95**	351.52
	swahili→embu	1044	**0.12**	15.00	**121.43**	0.04	**15.35**	215.73
	swahili→english	10844	0.46	15.97	183.94	**4.61**	**27.12**	**159.96**
	swahili→giriama	1885	**3.13**	**23.41**	**122.44**	0.70	17.28	243.88
	swahili→gusii	1269	**0.43**	15.85	**148.97**	0.16	**17.97**	306.31
	swahili→jita	1742	**0.34**	**14.25**	**136.45**	0.16	14.71	262.43
	swahili→kamba	1033	**0.11**	**15.27**	**123.54**	0.15	15.98	228.52
	swahili→kikuyu	1002	**0.14**	**14.73**	**126.40**	0.04	16.20	310.06
	swahili→luganda	1248	**0.06**	12.92	**115.26**	0.03	**14.92**	211.98
	swahili→luhya	2164	**0.13**	**12.86**	212.17	0.06	8.11	**151.67**
	swahili→luo	1577	**0.18**	**8.94**	**146.36**	0.14	10.18	307.79
	swahili→maasai	643	**0.18**	13.81	**157.32**	0.07	**15.57**	368.55
	swahili→meru	1044	**0.12**	**14.15**	**130.25**	0.04	14.82	277.16
	swahili→pokot	1044	**0.05**	**13.38**	**133.12**	0.01	15.09	323.54
English Paired Only	english→kalenjin	1431	**0.22**	**11.62**	**243.86**	0.29	10.91	807.11
	english→oromo	528	**0.17**	**9.50**	**130.13**	0.01	8.85	254.87
	english→samburu	1073	**0.02**	**5.56**	**653.14**	0.01	4.30	1952.75
	english→tigrinya	4	**0.00**	**0.00**	**1033.33**	0.12	0.52	1633.33
Swahili Paired Only	swahili→acholi	1248	**0.06**	12.47	**116.75**	0.02	**14.04**	245.16
	swahili→alur	1249	**0.11**	13.38	**120.66**	0.03	**14.48**	279.41
	swahili→aringa	1247	0.01	**9.60**	**110.17**	0.01	9.41	173.87
	swahili→ateso	1249	**0.08**	**11.75**	**116.56**	0.03	16.26	326.02
	swahili→bena	1731	**0.06**	**13.17**	**123.57**	0.02	13.85	242.49
	swahili→chaga	1949	**0.30**	13.83	**135.56**	0.10	**15.67**	297.54
	swahili→chonyi	73	**0.02**	**4.39**	**1927.03**	0.02	5.83	1863.51
	swahili→chuka	951	**0.12**	13.32	**131.91**	0.13	**15.55**	236.72
	swahili→duruma	1001	**0.61**	**16.45**	**117.76**	0.05	13.19	159.68
	swahili→gogo	6451	**0.29**	**13.50**	**113.66**	0.09	13.99	194.74
	swahili→gwere	1248	**0.04**	11.15	**120.21**	0.01	**15.93**	300.29
	swahili→haya	247	**0.38**	14.77	**213.70**	0.18	**15.18**	354.55
	swahili→hehe	1570	**0.24**	14.56	**122.05**	0.05	13.78	241.28
	swahili→jopadhola	1249	**0.08**	12.93	**124.53**	0.02	15.05	257.36
	swahili→kebu	1248	**0.07**	13.37	**122.33**	0.01	14.91	250.26
	swahili→kinyarwanda	1247	**0.06**	14.17	**144.35**	0.08	15.96	318.04
	swahili→kumam	1248	**0.05**	12.09	**111.17**	0.02	14.77	168.50
	swahili→langi	1568	**0.08**	14.57	**117.75**	0.02	15.31	228.27
	swahili→lango	1248	**0.05**	12.62	**115.40**	0.02	14.01	224.78
	swahili→luhya	2164	**0.13**	**12.86**	212.17	0.06	8.11	**151.67**
	swahili→masaba	1249	**0.06**	**13.39**	**125.37**	0.02	12.49	168.19
	swahili→meru	1044	**0.12**	**14.15**	**130.25**	0.04	14.82	277.16
	swahili→nyole	1249	**0.04**	12.82	**130.85**	0.01	14.87	237.88
	swahili→nyoro	1249	**0.13**	13.80	**168.38**	0.05	16.76	395.23
	swahili→pare	198	**0.02**	**2.33**	**173.44**	0.09	2.66	382.47
	swahili→pokomo	1001	**0.72**	**17.31**	**111.96**	0.24	21.54	270.14
	swahili→soga	1249	**0.04**	13.09	**144.90**	0.02	13.64	215.30
	swahili→somali	1000	**0.02**	12.44	**119.25**	0.04	15.31	236.68
	swahili→suba	1587	**0.14**	**13.44**	**121.25**	0.04	16.05	233.54
	swahili→sukuma	1277	**0.17**	13.80	**184.23**	0.07	14.29	320.88
	swahili→taita	5438	**0.30**	14.26	**188.38**	0.07	13.03	284.23
	swahili→taveta	1044	**0.59**	**17.92**	**143.53**	0.16	16.24	357.65
	swahili→zanaki	1561	**0.09**	12.11	**120.70**	0.02	15.62	290.97
Averaged Summary		66	**0.23**	11.48	**219.17**	0.15	**12.54**	413.99

Table 11: Zero-shot chrF++, BLEU, and TER scores for Gemma-3-270M and Gemma-3-1B base models (before any fine-tuning) on 66 East African language directions. For each direction and metric, the better score is in bold (higher is better for BLEU and chrF++; lower is better for TER).

Language	Country	Source Tokens	Translation Tokens	Fertility
High fertility (>1.5)				
Oromo	Ethiopia	77,261	160,014	**2.07**
Luganda	Uganda	3,079,301	5,061,186	**1.64**
Kikuyu	Kenya	2,141,717	3,510,370	**1.64**
Kalenjin	Kenya	84,952	135,554	**1.60**
Swahili	Kenya	2,668	4,200	**1.57**
Zanaki	Tanzania	321,367	485,813	**1.51**
Medium fertility (1.0–<1.5)				
Gwere	Uganda	231,740	340,491	**1.47**
Bena	Tanzania	334,552	480,999	**1.44**
Masaba	Uganda	231,741	328,161	**1.42**
Makonde	Tanzania	395,038	516,338	**1.31**
Gogo	Tanzania	1,456,656	1,865,915	**1.28**
Langi	Tanzania	322,965	409,292	**1.27**
Nyole	Uganda	231,740	294,609	**1.27**
Duruma	Kenya	197,413	248,717	**1.26**
Pare	Tanzania	15,001	18,819	**1.25**
Pokomo	Kenya	197,338	246,456	**1.25**
Kakwa	Uganda	231,740	289,423	**1.25**
Maasai	Tanzania	154,089	190,289	**1.24**
Suba	Kenya	361,159	437,423	**1.21**
Ateso	Uganda	231,740	280,299	**1.21**
Marakwet	Kenya	228,624	266,270	**1.16**
Soga	Uganda	231,740	267,777	**1.16**
Giriama	Kenya	327,412	376,830	**1.15**
Kebu	Uganda	231,740	265,804	**1.15**
Kumam	Uganda	231,740	265,015	**1.14**
Chaga	Tanzania	351,653	393,654	**1.12**
Hehe	Tanzania	325,353	365,460	**1.12**
Taita	Kenya	809,231	888,257	**1.10**
Haya	Tanzania	19,864	21,639	**1.09**
Jita	Tanzania	333,545	361,195	**1.08**
Luhya	Kenya	317,414	339,491	**1.07**
Borana	Kenya	247,267	264,151	**1.07**
Jopadhola	Uganda	231,740	247,727	**1.07**
Nyoro	Uganda	231,740	246,490	**1.06**
Acholi	Uganda	231,740	242,244	**1.05**
Kamba	Kenya	212,480	220,388	**1.04**
Lango	Uganda	231,741	237,077	**1.02**
Kinyarwanda	Rwanda	231,740	235,655	**1.02**
Low fertility (<1.0)				
Tigrinya	Ethiopia	408	90	**0.22**
English	Kenya	1,020,199	739,610	**0.73**
Taveta	Kenya	228,860	186,804	**0.82**
Pokot	Kenya	234,343	217,501	**0.93**
Chuka	Kenya	205,331	195,227	**0.95**
Chonyi	Kenya	1,359	1,291	**0.95**
Gusii	Kenya	269,601	261,580	**0.97**
Samburu	Kenya	47,140	45,595	**0.97**
Luo	Kenya	362,766	356,761	**0.98**
Aringa	Uganda	231,605	229,863	**0.99**
Embu	Kenya	228,901	226,948	**0.99**
Meru	Kenya	207,053	206,282	**1.00**

Table 12: Token-level fertility statistics across 53 East African languages, categorized into high (>1.5), medium (1.0–<1.5), and low fertility (<1.0). We compute fertility as a token-level ratio that measures how much a target-language translation expands or compresses the content of the source sentence.

Language	Pairs	Src Tok	Tgt Tok	Src Len	Tgt Len	Fert.
Luganda	162,446	3,079,301	5,061,186	18.96	31.16	1.64
Kikuyu	100,195	2,141,717	3,510,370	21.38	35.04	1.64
Taita	58,537	809,231	888,257	13.82	15.17	1.10
English	54,221	1,020,199	739,610	18.82	13.64	0.72
Gogo	32,371	1,456,656	1,865,915	45.00	57.64	1.28
Luhya	21,560	317,414	339,491	14.72	15.75	1.07
Chaga	10,106	351,653	393,654	34.79	38.96	1.12
Luo	10,038	362,766	356,761	36.12	35.55	0.98
Borana	9,697	247,267	264,151	25.50	27.25	1.07
Giriama	9,629	327,412	376,830	34.00	39.13	1.15
Makonde	9,111	395,038	516,338	43.37	56.67	1.31
Jita	8,820	333,545	361,195	37.81	40.96	1.08
Bena	8,674	334,552	480,999	38.56	55.44	1.44
Kamba	8,210	212,480	220,388	25.88	26.83	1.04
Kalenjin	7,969	84,952	135,554	10.66	17.01	1.60
Suba	7,965	361,159	437,423	45.33	54.91	1.21
Hehe	7,849	325,353	365,460	41.45	46.57	1.12
Langi	7,849	322,965	409,292	41.14	52.16	1.27
Zanaki	7,813	321,367	485,813	41.13	62.19	1.51
Pokot	6,533	234,343	217,501	35.87	33.30	0.93
Gusii	6,494	269,601	261,580	41.52	40.24	0.97
Sukuma	6,394	192,310	227,498	30.08	35.58	1.18
Masaba	6,246	231,741	328,161	37.10	52.54	1.41
Acholi	6,246	231,740	242,244	37.10	38.77	1.05
Alur	6,246	231,740	231,136	37.10	36.99	1.00
Ateso	6,246	231,740	280,299	37.10	44.88	1.21
Kakwa	6,246	231,740	289,423	37.10	46.32	1.25
Gwere	6,246	231,740	340,491	37.10	54.52	1.47
Jopadhola	6,246	231,740	247,727	37.10	39.66	1.07
Lango	6,246	231,741	237,077	37.10	37.96	1.02
Kebu	6,246	231,740	265,804	37.10	42.56	1.15
Kinyarwanda	6,246	231,740	235,655	37.10	37.73	1.02
Nyole	6,246	231,740	294,609	37.10	47.18	1.27
Nyoro	6,246	231,740	246,490	37.10	39.47	1.06
Soga	6,246	231,740	267,777	37.10	42.87	1.16
Kumam	6,246	231,740	265,015	37.10	42.43	1.14
Aringa	6,243	231,605	229,863	37.09	36.80	0.99
Meru	5,420	207,053	206,282	38.21	38.06	1.00
Samburu	5,411	47,140	45,595	8.71	8.43	0.97
Embu	5,349	228,901	226,948	42.79	42.43	0.99
Maasai	5,327	154,089	190,289	28.93	35.73	1.23
Taveta	5,249	228,860	186,804	43.62	35.60	0.82
Marakwet	5,249	228,624	266,270	43.57	50.75	1.16
Chuka	5,030	205,331	195,227	40.82	38.81	0.95
Duruma	5,011	197,413	248,717	39.40	49.64	1.26
Somali	5,011	197,413	235,446	39.40	46.97	1.19
Pokomo	5,010	197,338	246,456	39.40	49.17	1.25
Oromo	3,423	77,261	160,014	22.57	46.75	2.07
Haya	1,306	19,864	21,639	15.21	16.57	1.09
Pare	1,000	15,001	18,819	15.00	18.82	1.26
Chonyi	370	1,359	1,291	3.67	3.49	0.95
Swahili	138	2,668	4,200	19.34	30.43	1.57
Tigrinya	18	408	90	22.67	5.00	0.22

Table 13: Sequence length and fertility statistics for all languages.

Lang Pair	#samples	Llama-3-1B				Qwen-3-0.6B				Qwen-3-1.7B			
		X→Y		Y→X		X→Y		Y→X		X→Y		Y→X	
		BLEU	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU	ChrF++	BLEU	ChrF++
acholi↔swahili	1248	3.43	14.36	4.93	12.23	0.67	12.81	1.57	12.15	0.63	12.86	1.22	12.30
alur↔swahili	1249	11.54	15.40	6.92	13.60	0.57	13.46	1.46	12.84	0.61	13.77	0.79	12.08
aringa↔swahili	1247	1.78	14.45	1.69	8.50	0.49	11.78	0.79	7.89	0.55	11.91	0.76	7.80
ateso↔swahili	1249	1.86	15.33	4.46	13.18	0.57	13.26	0.64	14.11	0.56	13.55	0.65	14.43
bena↔swahili	1731	6.34	16.85	3.96	14.70	0.58	13.01	1.36	14.61	1.42	14.45	1.26	14.98
borana↔english	889	4.07	4.43	1.38	5.53	0.65	2.52	0.65	3.12	0.64	2.50	0.65	3.10
borana↔swahili	1044	20.95	15.83	3.74	14.15	0.75	13.89	0.76	13.75	0.70	14.11	0.65	13.69
chaga↔swahili	1949	6.64	14.93	3.15	16.24	3.74	14.49	1.51	14.35	4.40	15.04	1.78	14.62
chonyi↔swahili	73	0.00	3.67	0.00	5.62	0.33	1.94	0.00	1.94	0.34	2.04	0.00	1.99
chuka↔swahili	951	4.92	15.71	2.17	14.45	1.05	14.20	0.85	13.54	1.72	14.24	0.92	13.83
duruma↔swahili	1001	3.03	20.24	3.85	18.52	0.58	16.45	1.58	17.89	1.57	17.51	2.14	18.41
embu↔english	20	0.00	3.80	0.00	4.17	0.00	1.48	0.30	1.02	0.34	1.55	0.00	1.11
embu↔swahili	1044	4.71	16.87	4.62	14.38	0.77	14.48	1.66	14.30	0.77	14.68	0.79	14.76
english↔giriama	26	7.81	11.93	2.65	17.16	0.53	10.18	1.67	14.76	0.54	9.88	0.73	14.92
english↔gusii	20	0.00	5.00	1.26	2.61	0.00	1.75	0.00	1.39	0.00	1.72	0.00	1.46
english↔jita	21	0.29	4.67	3.46	7.42	0.30	2.21	0.43	2.93	0.28	2.30	0.43	3.21
english↔kalenjin	1431	10.68	11.01	5.44	12.75	1.12	8.77	2.85	11.27	1.51	8.95	3.00	11.12
english↔kamba	180	0.00	3.61	2.95	3.86	0.30	1.42	0.63	1.90	0.32	1.46	1.30	2.00
english↔kikuyu	18989	5.24	13.65	26.27	15.10	2.04	12.04	11.57	14.59	3.97	11.97	17.55	14.83
english↔luganda	23875	5.02	11.51	6.04	15.53	1.25	9.50	4.57	13.29	1.21	9.77	6.30	13.30
english↔luhya	1518	3.67	5.44	2.21	5.97	0.52	2.03	2.27	4.23	0.53	2.05	2.40	4.28
english↔luo	154	1.31	12.98	3.09	14.67	0.86	8.65	2.09	11.67	0.93	8.94	2.17	11.81
english↔maasai	416	1.22	14.34	4.25	15.46	1.36	12.42	5.73	15.66	1.30	12.37	4.18	15.58
english↔meru	39	15.97	4.58	2.44	5.98	0.35	1.99	0.50	2.32	0.35	2.02	0.46	2.54
english↔oromo	528	3.67	10.97	6.08	15.72	0.55	10.84	2.48	16.04	0.56	10.83	2.46	15.62
english↔pokot	260	0.00	6.81	3.05	5.99	0.39	2.52	1.44	3.82	0.39	2.50	1.36	3.99
english↔samburu	1073	0.00	5.18	3.98	8.85	0.54	2.32	2.34	6.53	0.55	2.31	3.52	6.57
english↔swahili	10870	14.38	22.30	5.52	31.34	1.23	9.09	5.26	12.47	1.42	9.67	9.67	13.08
english↔tigrinya	4	2.28	4.58	0.89	9.36	0.48	2.25	1.40	8.56	0.39	2.17	0.79	7.44
giriama↔swahili	1885	8.89	29.38	11.02	25.43	1.25	20.04	1.93	22.11	1.15	22.56	1.36	23.22
gogo↔swahili	6451	3.30	16.12	3.55	14.52	0.64	13.48	1.67	13.66	0.73	14.23	2.11	14.42
gusii↔swahili	1269	3.00	17.15	7.27	14.62	0.72	15.03	1.12	15.60	0.77	14.96	1.28	15.81
gwere↔swahili	1248	4.27	15.92	4.93	13.12	0.56	13.53	0.65	14.63	0.66	13.91	0.67	14.82
haya↔swahili	247	4.20	20.96	5.52	21.53	0.47	10.60	0.53	12.90	0.52	12.98	0.50	12.87
hehe↔swahili	1570	3.21	17.15	3.46	15.43	3.76	13.94	0.73	14.36	3.62	15.19	0.77	15.06
jita↔swahili	1742	4.13	17.38	3.21	16.42	1.37	15.03	1.42	14.63	1.33	15.43	0.86	14.95
jopadhola↔swahili	1249	1.00	13.40	2.02	12.21	0.56	13.11	1.54	12.92	0.75	13.20	1.49	13.10
kakwa↔swahili	1249	2.65	15.70	7.77	12.88	0.67	13.15	1.69	14.11	0.66	13.46	1.62	13.81
kamba↔swahili	1033	4.62	16.63	2.73	15.97	1.20	14.70	1.33	14.07	1.13	14.85	0.94	14.53
kebu↔swahili	1248	2.63	14.99	2.40	13.17	0.68	13.37	0.60	14.11	0.74	13.34	0.90	14.19
kikuyu↔swahili	1002	5.21	16.23	2.35	14.75	0.67	13.60	0.82	14.26	0.61	14.34	0.83	14.20
kinyarwanda↔swahili	1247	4.14	15.86	6.81	14.71	1.64	14.08	1.74	13.33	1.55	14.42	1.70	13.53
kumam↔swahili	1248	2.30	16.10	3.04	11.93	0.95	12.83	0.74	13.42	0.97	12.96	0.89	13.48
langi↔swahili	1568	2.16	18.33	3.66	15.37	0.75	14.89	1.47	15.54	0.85	15.52	1.81	14.80
lango↔swahili	1248	2.60	15.45	3.42	11.66	0.56	12.61	1.58	12.34	0.64	12.78	1.27	12.39
luganda↔swahili	1248	1.61	15.05	3.28	13.20	1.15	13.15	0.78	12.97	1.20	13.64	1.22	13.15
luhya↔swahili	2164	3.06	15.49	7.47	14.63	1.34	10.68	1.47	11.15	1.28	10.99	1.31	11.15
luo↔swahili	1577	2.71	13.35	2.86	11.36	0.69	12.78	1.60	10.62	1.15	12.94	1.96	10.98
maasai↔swahili	643	1.81	16.28	1.45	14.25	0.69	13.61	0.81	14.08	0.74	13.63	0.80	14.02
makonde↔swahili	1820	8.58	19.78	5.13	17.50	0.90	16.26	0.67	16.97	0.86	16.21	1.62	17.91
marakwet↔swahili	1044	4.93	13.98	4.07	11.89	0.73	13.38	0.76	12.49	0.62	12.98	0.65	12.67
masaba↔swahili	1249	2.17	16.25	1.93	13.95	0.64	13.35	1.37	15.11	0.71	14.03	1.33	15.25
meru↔swahili	1044	2.78	15.61	3.74	13.71	0.70	13.15	1.52	13.81	0.75	13.96	0.84	13.95
nyole↔swahili	1249	1.19	14.99	2.26	12.74	0.72	12.90	0.74	13.81	0.74	13.13	0.78	13.85
nyoro↔swahili	1249	4.30	16.39	4.30	14.94	0.79	14.01	1.40	14.17	0.82	14.56	1.41	14.51
oromo↔english	528	6.08	15.72	3.67	10.97	2.48	16.04	0.55	10.84	2.46	15.62	0.56	10.83
pare↔swahili	198	3.75	9.59	2.35	3.49	0.36	8.77	0.56	2.61	0.38	8.91	0.62	2.40
pokomo↔swahili	1001	3.88	22.18	10.10	20.34	0.81	16.60	1.57	18.34	0.74	19.41	1.59	19.60
pokot↔swahili	1044	4.42	14.53	1.10	12.90	0.66	12.89	1.54	12.93	0.74	13.63	1.64	13.01
samburu↔english	1073	3.98	8.85	0.00	5.18	2.34	6.53	0.54	2.32	3.52	6.57	0.55	2.31
soga↔swahili	1249	1.83	14.98	3.80	14.26	0.77	13.84	1.36	13.81	0.65	13.94	1.43	13.98
somali↔swahili	1000	1.82	15.61	3.46	12.47	0.63	13.10	1.60	13.36	0.75	13.28	2.03	13.28
suba↔swahili	1587	13.72	16.98	1.67	15.17	0.65	14.99	0.86	15.22	0.72	15.41	0.88	15.73
sukuma↔swahili	1277	4.02	15.41	4.25	15.70	0.58	13.27	1.18	13.64	0.66	14.01	1.18	13.86
swahili↔taita	5438	10.68	17.02	13.61	16.84	0.62	10.05	0.66	10.39	1.18	10.15	1.28	10.50
swahili↔taveta	1044	2.68	18.21	4.56	18.50	1.48	15.40	1.52	15.89	1.48	15.84	1.48	16.58
tigrinya↔english	4	0.89	9.36	2.28	4.58	1.40	8.56	0.48	2.25	0.79	7.44	0.39	2.17
zanaki↔swahili	1561	12.65	16.56	5.04	13.19	1.18	13.93	1.42	15.19	0.64	14.59	1.52	15.39

Table 14: Bidirectional BLEU and ChrF++ scores for **Llama-3-1B**, **Qwen-3-0.6B**, and **Qwen-3-1.7B** on AfriMMT-EA language pairs. X→Y and Y→X denote source and reverse translation directions for each language pair.

Lang Pair	#samples	$L_1 \rightarrow L_2$		$L_2 \rightarrow L_1$	
		BLEU	ChrF++	BLEU	ChrF++
acholi↔swahili	1248	3.43	14.36	4.93	12.23
alur↔swahili	1249	11.54	15.4	6.92	13.6
aringa↔swahili	1247	1.78	14.45	1.69	8.5
ateso↔swahili	1249	1.86	15.33	4.46	13.18
bena↔swahili	1731	6.34	16.85	3.96	14.7
borana↔english	889	4.07	4.43	1.38	5.53
borana↔swahili	1044	20.95	15.83	3.74	14.15
chaga↔swahili	1949	6.64	14.93	3.15	16.24
chonyi↔swahili	73	0	3.67	0	5.62
chuka↔swahili	951	4.92	15.71	2.17	14.45
duruma↔swahili	1001	3.03	20.24	3.85	18.52
embu↔english	20	0	3.8	0	4.17
embu↔swahili	1044	4.71	16.87	4.62	14.38
english↔giriama	26	7.81	11.93	2.65	17.16
english↔gusii	20	0	5	1.26	2.61
english↔jita	21	0.29	4.67	3.46	7.42
english↔kalenjin	1431	10.68	11.01	5.44	12.75
english↔kamba	180	0	3.61	2.95	3.86
english↔kikuyu	18989	5.24	13.65	26.27	15.1
english↔luganda	23875	5.02	11.51	6.04	15.53
english↔luhya	1518	3.67	5.44	2.21	5.97
english↔luo	154	1.31	12.98	3.09	14.67
english↔maasai	416	1.22	14.34	4.25	15.46
english↔meru	39	15.97	4.58	2.44	5.98
english↔oromo	528	3.67	10.97	6.08	15.72
english↔pokot	260	0	6.81	3.05	5.99
english↔samburu	1073	0	5.18	3.98	8.85
english↔swahili	10870	14.38	22.3	5.52	31.34
english↔tigrinya	4	2.28	4.58	0.89	9.36
giriama↔swahili	1885	8.89	29.38	11.02	25.43
gogo↔swahili	6451	3.3	16.12	3.55	14.52
gusii↔swahili	1269	3	17.15	7.27	14.62
gwere↔swahili	1248	4.27	15.92	4.93	13.12
haya↔swahili	247	4.2	20.96	5.52	21.53
hehe↔swahili	1570	3.21	17.15	3.46	15.43
jita↔swahili	1742	4.13	17.38	3.21	16.42
jopadhola↔swahili	1249	1	13.4	2.02	12.21
kakwa↔swahili	1249	2.65	15.7	7.77	12.88
kalenjin↔swahili	1431				
kamba↔swahili	1033	4.62	16.63	2.73	15.97
kebu↔swahili	1248	2.63	14.99	2.4	13.17
kikuyu↔swahili	1002	5.21	16.23	2.35	14.75
kinyarwanda↔swahili	1247	4.14	15.86	6.81	14.71
kumam↔swahili	1248	2.3	16.1	3.04	11.93
langi↔swahili	1568	2.16	18.33	3.66	15.37
lango↔swahili	1248	2.6	15.45	3.42	11.66
luganda↔swahili	1248	1.61	15.05	3.28	13.2
luhya↔swahili	2164	3.06	15.49	7.47	14.63
luo↔swahili	1577	2.71	13.35	2.86	11.36
maasai↔swahili	643	1.81	16.28	1.45	14.25
makonde↔swahili	1820	8.58	19.78	5.13	17.5
marakwet↔swahili	1044	4.93	13.98	4.07	11.89
masaba↔swahili	1249	2.17	16.25	1.93	13.95
meru↔swahili	1044	2.78	15.61	3.74	13.71
nyole↔swahili	1249	1.19	14.99	2.26	12.74
nyoro↔swahili	1249	4.3	16.39	4.3	14.94
oromo↔swahili	528				
pare↔swahili	198	3.75	9.59	2.35	3.49
pokomo↔swahili	1001	3.88	22.18	10.1	20.34
pokot↔swahili	1044	4.42	14.53	1.1	12.9
samburu↔swahili	1073				
soga↔swahili	1249	1.83	14.98	3.8	14.26
somali↔swahili	1000	1.82	15.61	3.46	12.47
suba↔swahili	1587	13.72	16.98	1.67	15.17
sukuma↔swahili	1277	4.02	15.41	4.25	15.7
swahili↔taita	5438	10.68	17.02	13.61	16.84
swahili↔taveta	1044	2.68	18.21	4.56	18.5
swahili↔zanaki	1561	5.04	13.19	12.65	16.56
tigrinya↔english	4	0.89	9.36	2.28	4.58

Table 15: BLEU and ChrF++ for **Llama-3-1B** on bidirectional AfriMMT-EA language pairs. $L_1 \rightarrow L_2$ and $L_2 \rightarrow L_1$ share the same #samples within each pair.

Lang Pair	#samples	$L_1 \rightarrow L_2$		$L_2 \rightarrow L_1$	
		BLEU	ChrF++	BLEU	ChrF++
acholi↔swahili	1248	0.67	12.81	1.57	12.15
alur↔swahili	1249	0.57	13.46	1.46	12.84
aringa↔swahili	1247	0.49	11.78	0.79	7.89
ateso↔swahili	1249	0.57	13.26	0.64	14.11
bena↔swahili	1731	0.58	13.01	1.36	14.61
borana↔english	889	0.65	2.52	0.65	3.12
borana↔swahili	1044	0.75	13.89	0.76	13.75
chaga↔swahili	1949	3.74	14.49	1.51	14.35
chonyi↔swahili	73	0.33	1.94	0	1.94
chuka↔swahili	951	1.05	14.20	0.85	13.54
duruma↔swahili	1001	0.58	16.45	1.58	17.89
embu↔english	20	0	1.48	0.30	1.02
embu↔swahili	1044	0.77	14.48	1.66	14.30
english↔giriama	26	0.53	10.18	1.67	14.76
english↔gusii	20	0	1.75	0	1.39
english↔jita	21	0.30	2.21	0.43	2.93
english↔kalenjin	1431	1.12	8.77	2.85	11.27
english↔kamba	180	0.30	1.42	0.63	1.90
english↔kikuyu	18989	2.04	12.04	11.57	14.59
english↔luganda	23875	1.25	9.50	4.57	13.29
english↔luhya	1518	0.52	2.03	2.27	4.23
english↔luo	154	0.86	8.65	2.09	11.67
english↔maasai	416	1.36	12.42	5.73	15.66
english↔meru	39	0.35	1.99	0.50	2.32
english↔oromo	528	0.55	10.84	2.48	16.04
english↔pokot	260	0.39	2.52	1.44	3.82
english↔samburu	1073	0.54	2.32	2.34	6.53
english↔swahili	10870	1.23	9.09	5.26	12.47
english↔tigrinya	4	0.48	2.25	1.40	8.56
giriama↔swahili	1885	1.25	20.04	1.93	22.11
gogo↔swahili	6451	0.64	13.48	1.67	13.66
gusii↔swahili	1269	0.72	15.03	1.12	15.60
gwere↔swahili	1248	0.56	13.53	0.65	14.63
haya↔swahili	247	0.47	10.60	0.53	12.90
hehe↔swahili	1570	3.76	13.94	0.73	14.36
jita↔swahili	1742	1.37	15.03	1.42	14.63
jopadhola↔swahili	1249	0.56	13.11	1.54	12.92
kakwa↔swahili	1249	0.67	13.15	1.69	14.11
kalenjin↔swahili	1431				
kamba↔swahili	1033	1.20	14.70	1.33	14.07
kebu↔swahili	1248	0.68	13.37	0.60	14.11
kikuyu↔swahili	1002	0.67	13.60	0.82	14.26
kinyarwanda↔swahili	1247	1.64	14.08	1.74	13.33
kumam↔swahili	1248	0.95	12.83	0.74	13.42
langi↔swahili	1568	0.75	14.89	1.47	15.54
lango↔swahili	1248	0.56	12.61	1.58	12.34
luganda↔swahili	1248	1.15	13.15	0.78	12.97
luhya↔swahili	2164	1.34	10.68	1.47	11.15
luo↔swahili	1577	0.69	12.78	1.60	10.62
maasai↔swahili	643	0.69	13.61	0.81	14.08
makonde↔swahili	1820	0.90	16.26	0.67	16.97
marakwet↔swahili	1044	0.73	13.38	0.76	12.49
masaba↔swahili	1249	0.64	13.35	1.37	15.11
meru↔swahili	1044	0.70	13.15	1.52	13.81
nyole↔swahili	1249	0.72	12.90	0.74	13.81
nyoro↔swahili	1249	0.79	14.01	1.40	14.17
oromo↔swahili	528				
pare↔swahili	198	0.36	8.77	0.56	2.61
pokomo↔swahili	1001	0.81	16.60	1.57	18.34
pokot↔swahili	1044	0.66	12.89	1.54	12.93
samburu↔swahili	1073				
soga↔swahili	1249	0.77	13.84	1.36	13.81
somali↔swahili	1000	0.63	13.10	1.60	13.36
suba↔swahili	1587	0.65	14.99	0.86	15.22
sukuma↔swahili	1277	0.58	13.27	1.18	13.64
swahili↔taita	5438	0.62	10.05	0.66	10.39
swahili↔taveta	1044	1.48	15.40	1.52	15.89
swahili↔zanaki	1561	1.42	15.19	1.18	13.93
tigrinya↔english	4	1.40	8.56	0.48	2.25

Table 16: BLEU and ChrF++ for **Qwen-3-0.6B** on bidirectional AfriMMT-EA language pairs.

Lang Pair	#samples	$L_1 \rightarrow L_2$		$L_2 \rightarrow L_1$	
		BLEU	ChrF++	BLEU	ChrF++
acholi↔swahili	1248	0.63	12.86	1.22	12.30
alur↔swahili	1249	0.61	13.77	0.79	12.08
aringa↔swahili	1247	0.55	11.91	0.76	7.80
ateso↔swahili	1249	0.56	13.55	0.65	14.43
bena↔swahili	1731	1.42	14.45	1.26	14.98
borana↔english	889	0.64	2.50	0.65	3.10
borana↔swahili	1044	0.70	14.11	0.65	13.69
chaga↔swahili	1949	4.40	15.04	1.78	14.62
chonyi↔swahili	73	0.34	2.04	0	1.99
chuka↔swahili	951	1.72	14.24	0.92	13.83
duruma↔swahili	1001	1.57	17.51	2.14	18.41
embu↔english	20	0.34	1.55	0	1.11
embu↔swahili	1044	0.77	14.68	0.79	14.76
english↔giriama	26	0.54	9.88	0.73	14.92
english↔gusii	20	0	1.72	0	1.46
english↔jita	21	0.28	2.30	0.43	3.21
english↔kalenjin	1431	1.51	8.95	3.00	11.12
english↔kamba	180	0.32	1.46	1.30	2.00
english↔kikuyu	18989	3.97	11.97	17.55	14.83
english↔luganda	23875	1.21	9.77	6.30	13.30
english↔luhya	1518	0.53	2.05	2.40	4.28
english↔luo	154	0.93	8.94	2.17	11.81
english↔maasai	416	1.30	12.37	4.18	15.58
english↔meru	39	0.35	2.02	0.46	2.54
english↔oromo	528	0.56	10.83	2.46	15.62
english↔pokot	260	0.39	2.50	1.36	3.99
english↔samburu	1073	0.55	2.31	3.52	6.57
english↔swahili	10870	1.42	9.67	9.67	13.08
english↔tigrinya	4	0.39	2.17	0.79	7.44
giriama↔swahili	1885	1.15	22.56	1.36	23.22
gogo↔swahili	6451	0.73	14.23	2.11	14.42
gusii↔swahili	1269	0.77	14.96	1.28	15.81
gwere↔swahili	1248	0.66	13.91	0.67	14.82
haya↔swahili	247	0.52	12.98	0.50	12.87
hehe↔swahili	1570	3.62	15.19	0.77	15.06
jita↔swahili	1742	1.33	15.43	0.86	14.95
jopadhola↔swahili	1249	0.75	13.20	1.49	13.10
kakwa↔swahili	1249	0.66	13.46	1.62	13.81
kalenjin↔swahili	1431				
kamba↔swahili	1033	1.13	14.85	0.94	14.53
kebu↔swahili	1248	0.74	13.34	0.90	14.19
kikuyu↔swahili	1002	0.61	14.34	0.83	14.20
kinyarwanda↔swahili	1247	1.55	14.42	1.70	13.53
kumam↔swahili	1248	0.97	12.96	0.89	13.48
langi↔swahili	1568	0.85	15.52	1.81	14.80
lango↔swahili	1248	0.64	12.78	1.27	12.39
luganda↔swahili	1248	1.20	13.64	1.22	13.15
luhya↔swahili	2164	1.28	10.99	1.31	11.15
luo↔swahili	1577	1.15	12.94	1.96	10.98
maasai↔swahili	643	0.74	13.63	0.80	14.02
makonde↔swahili	1820	0.86	16.21	1.62	17.91
marakwet↔swahili	1044	0.62	12.98	0.65	12.67
masaba↔swahili	1249	0.71	14.03	1.33	15.25
meru↔swahili	1044	0.75	13.96	0.84	13.95
nyole↔swahili	1249	0.74	13.13	0.78	13.85
nyoro↔swahili	1249	0.82	14.56	1.41	14.51
oromo↔swahili	528				
pare↔swahili	198	0.38	8.91	0.62	2.40
pokomo↔swahili	1001	0.74	19.41	1.59	19.60
pokot↔swahili	1044	0.74	13.63	1.64	13.01
samburu↔swahili	1073				
soga↔swahili	1249	0.65	13.94	1.43	13.98
somali↔swahili	1000	0.75	13.28	2.03	13.28
suba↔swahili	1587	0.72	15.41	0.88	15.73
sukuma↔swahili	1277	0.66	14.01	1.18	13.86
swahili↔taita	5438	1.18	10.15	1.28	10.50
swahili↔taveta	1044	1.48	15.84	1.48	16.58
swahili↔zanaki	1561	1.52	15.39	0.64	14.59
tigrinya↔english	4	0.79	7.44	0.39	2.17

Table 17: BLEU and ChrF++ for **Qwen-3-1.7B** on bidirectional AfriMMT-EA language pairs.

Translation Annotation Guidelines

Task Description
Volunteers translate text between **Swahili** (source) and **Kihehe** (target).

Instructions

1. Read the Swahili source text carefully.
2. Provide an accurate and complete translation in Kihehe.
3. Click outside the translation box or press `Tab` to save your entry.
4. Rate the translation quality using the provided scale.

Important Notes

- Your translation syncs automatically with the **Final Translation** field.
- The sync indicator shows if both fields match.
- If you edit the Final Translation directly, click `Sync` to update your translation.

Quality Guidelines

- Preserve the meaning of the source text.
- Use natural, fluent Kihehe.
- Maintain tone and style when appropriate.
- Keep terminology consistent across translations.

Figure 12: Sample Volunteer Translation Annotation Guidelines for Swahili–Kihehe. These guidelines are found inside Argilla web-based interface shown in Figure 7